COVID-19 Vulnerability in America

Inequality in Risk and Exposure to a Global Pandemic

Christopher Delaney, Austin Eaton, Clara Richter, Elise Rust

ANLY 511 Final Project

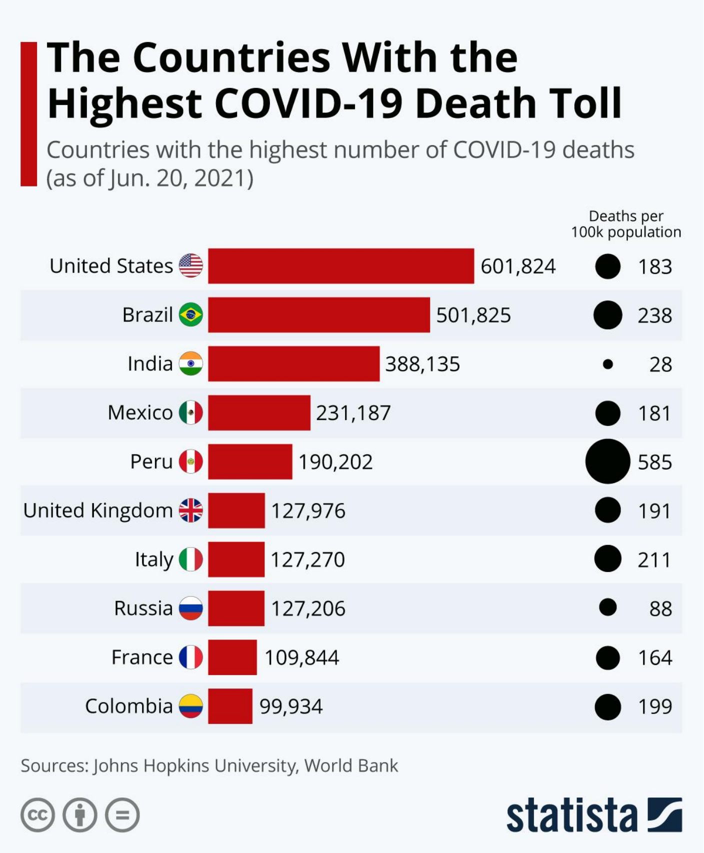
Georgetown University

8 December 2021

I. Introduction

The year 2020 rocked global society and the economy as the novel Coronavirus spread rapidly throughout the world. Citizens were ordered to quarantine at home indefinitely as public transit came to a halt, company offices and businesses closed their doors, and hospital systems began to face overwhelming numbers of patients. Upon publication of this report the world is coming up on the two year anniversary of COVID-19’s entrance on the international stage and the pandemic has been a life defining moment that has affected how society functions on both an individual and institutional basis. The United States of America has had an especially difficult time containing the spread of the virus and has seen some of the worst devastation on a per capita basis than most other countries.

Figure 1



Sources: Johns Hopkins University, World Bank

According to The New York Times, when comparing all nations, the United States ranks first in both the number of total reported cases and the number of total deaths due to the virus (NYT 2020). A myriad of factors - among them the politicization of mask wearing and vaccine hesitancy - have contributed to this and have led to over 50 million cases and 800,000 deaths from COVID-19 in the U.S.

Thus, the health risks of COVID-19 remain prominent. For those that survive an infection with the virus, many new reports are coming out about long term consequences on brain, lung, and heart health. Many people, even those who are young, have been known to describe a “fog” that lingers after COVID-19 infection that makes work and productivity much more difficult. Given these health consequences, being able to identify trends and patterns in who is most vulnerable to contracting COVID-19 is powerful.

COVID-19 cases are not uniform throughout the country, and many regions have experienced higher rates of the virus than others. Identifying these at-risk locations can help direct more resources and information towards them to best mitigate the risk and prevent the most number of people from suffering from this virus’ consequences. Therefore, the main questions this report seeks to address are:

1. *Do people in America face different risk profiles in contracting COVID-19 based on their location?*
2. *Are some subsets of the American population more susceptible? Are some subsets better equipped to avoid contracting COVID-19?*
3. *What factors influence this risk?*

The key risk factors examined are **population size, mask use, and political leanings** of population subgroups. This is examined at both the county and state level, as well as in high risk epicenters such as prisons and college campuses nationwide.

II. About the Data

A. Data Sources

This analysis primarily uses The New York Times Coronavirus (COVID-19) data to answer these questions. This dataset contains four tables of cumulative counts of coronavirus cases in the United States, at the state and county level, as well as at the prison and college level from the start of the pandemic until November 1st, 2021. Individuals included in the data are counted where they are being treated/tested for COVID-19, not necessarily where they live.

The college dataset was constructed from a survey between July 2020 and May 2020. The survey used reports from students and employees from every four-year public institution and every private college that competes in N.C.A.A. sports.

The mask-use-by-county data came from a large number of interviews that asked "How often do you wear a mask in public when you expect to be within six feet of another person?" The interviews were gathered once in early June 2020 and obtained 250,000 responses.

Data for the Facilities dataset was collected from March 2020 to March 2021 on cases and deaths in state and federal prisons; immigration detention centers; juvenile detention facilities; local, regional and reservation jails; and those in the custody of the U.S. Marshals Service. The data was collected from websites overseen by state and federal prison systems and Immigration and Customs Enforcement or through direct inquiries and public records requests, and from data presented at news conferences and meetings of county or state officials.

College enrollment data in 2020 was gathered from private company StateUniversity.com who provide the only publicly available enrollment data for the top 500 ranked colleges across the nation. Political affiliation data was provided by MIT's Election Data Science Lab which contained county-level results for presidential elections from 2000 to 2020. Additionally, a dataset with total population by county and a dataset with number of votes cast for each party in the 2020 election were acquired from the US Census Bureau.

B. Data Cleaning + Exploratory Data Analysis

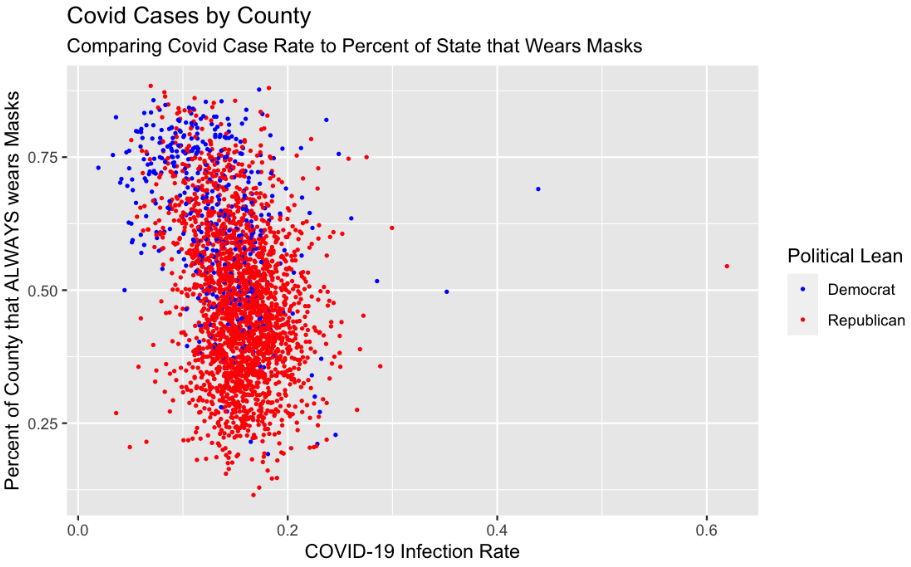
The college dataset was fairly well maintained but required substantial merging to combine college case data with enrollment numbers in 2020, mask use data in the surrounding counties, and election results of the surrounding counties. After initial data type correction and filtering for election results in 2020 only, NA values were addressed via median imputation. Median imputation was chosen as an alternative to mean imputation to preserve more of the data's integrity without losing observations so as to analyze as many colleges as possible. After cleaning, college COVID-19 case data was joined with enrollment and mask usage data by "county name" and "state". Case rate was again calculated by dividing total number of cases by student enrollment numbers in 2020. The prison dataset underwent a very similar cleaning process in order to join facility data such as inmate population, case rates, and facility type with external data such as surrounding county mask usage and political leaning.

To evaluate differing covid case and death rates between groups of counties based on frequency of masking, the mask dataset, county population dataset, as well as the covid case and death numbers were utilized. For each dataset, the column name corresponding to the fips county code was named "fips" and transformed to a numeric value if necessary in order to merge the three datasets. Once merged, the covid case rate was calculated by dividing the number of confirmed cases obtained from the covid dataset by the total population from each county given by the Census Bureau dataset. The same process was used to calculate the covid death rate. The rows were then arranged from highest to lowest for the proportion of people who answered “never” in response to how often they wore masks, counties with cases or deaths listed as NA were dropped to avoid missing values, and the top 30 resulting counties were selected and input into a separate dataset to represent a subset of counties with low mask usage. The same process was repeated, except with counties who answered “always” in response to how often they wore masks, to find the 30 counties with the highest mask usage. These two separate datasets were then combined into a final dataset, with the low mask counties and high mask counties labeled accordingly, along with their respective corresponding covid case and death rates. Lastly, we merged the counties dataset with datasets containing information regarding how the county voted in the 2020 presidential election, as well as information regarding how the state the county is located in voted in the 2020 presidential election.

Exploratory Data Analysis

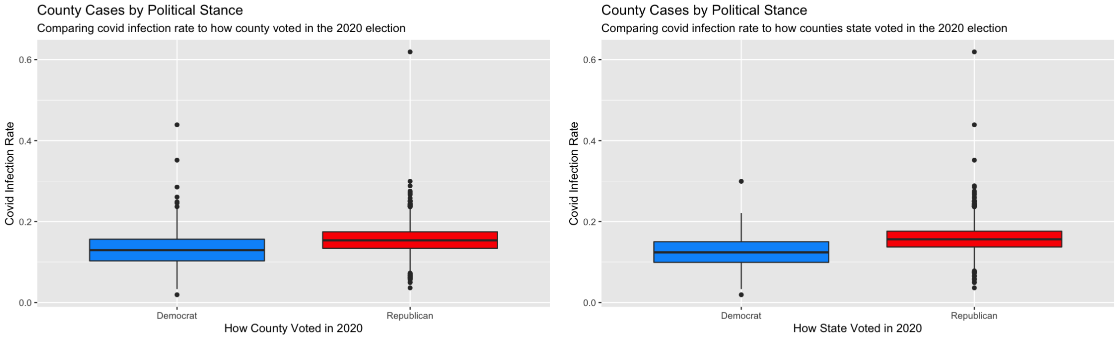
Initial exploratory data analysis offers interesting insight into patterns between these variables. Figure 2 highlights COVID-19 infection rates versus the percentage of a county that always wears masks, showing that most counties with a high mask usage rate also voted Democratic in the 2020 election. However, this plot doesn’t necessarily indicate a strong linear relationship between infection mask usage and infection rate.

Figure 2



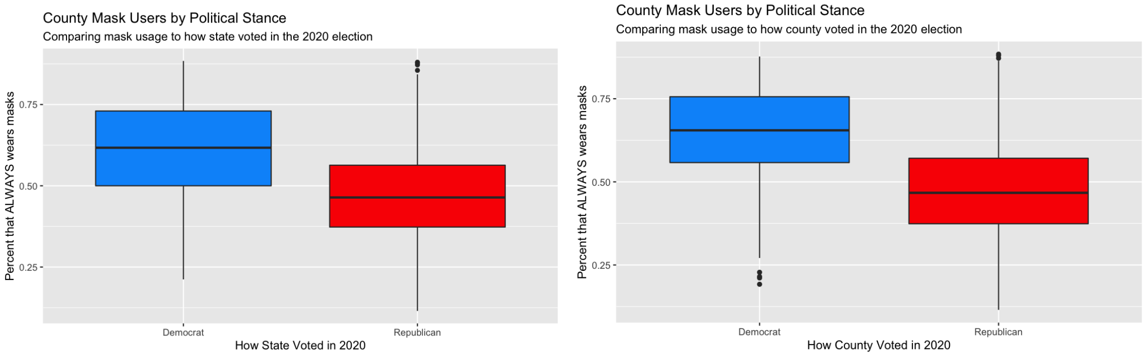
A series of boxplots examine the difference in mean case rates across different subgroups. For counties and states that voted Republican in the 2020 election, the average COVID-19 infection rate is slightly higher than counties and states that voted Democratic. Republican counties also tended to have a wider range of infection rates than Democratic counties, which raises questions about the socioeconomic and demographic composition of the two groups of counties.

Figure 3



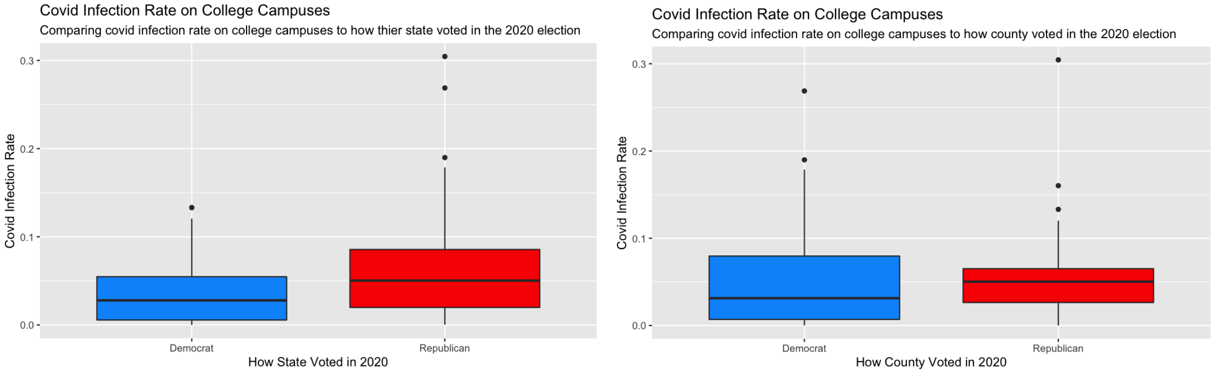
For counties and states that voted Republican in the 2020 election, the mask usage rate is lower on average than for counties and states that voted Democratic. A median of 60% of counties that voted Democrat reported always wearing masks, while a median of 49% of counties that voted Republican reported always wearing masks.

Figure 4



College Level Relationships

Figure 5



Similar relationships were visualized at the college level. Figure 5 illustrates COVID-19 infection rate against political leaning of the surrounding county using a college's COVID-19 rate. The same trend holds of colleges in Republican voting counties seeing higher infection rates on average than colleges in Democratic counties. It is interesting to compare these rates to those of the surrounding county to pinpoint how similar a college's infection rate is to their county's infection rate. Are the campus boundaries fairly porous? Is there significant overlap between the college population and the county population that would lead to high transmission rates between the groups? Or is the college fairly self contained? From these plots it appears that on average a campuses infection rates mirror those of its surrounding county, but the variance on campuses is larger. This makes sense given that college's make separate COVID-19 prevention plans than their local government.

Figure 6

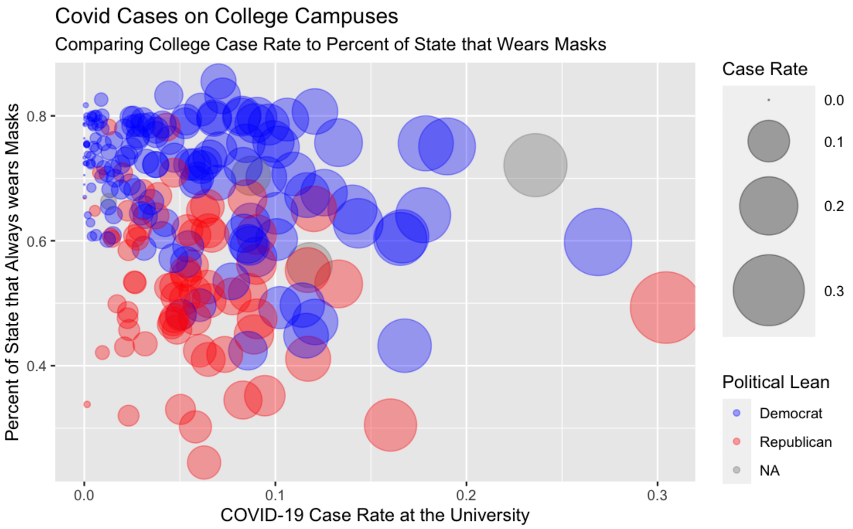
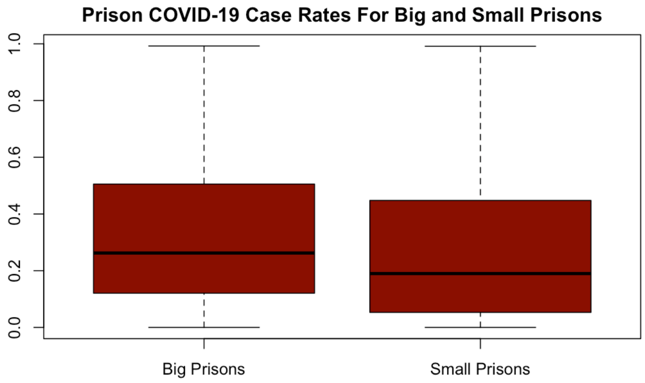


Figure 6 examines many of these relationships simultaneously by graphically depicting COVID-19 case rates in colleges against the percent of the state that always wears a mask. Counties are colored by their political results in the 2020 election and size of the point varies by case rate. As seen, many Democratic counties (in blue) exist in the upper left corner of the plot where cases are low and mask usage is high - supporting the hypothesis that Democratic counties tend to have higher rates of mask usage than Republican counties. This plot also offers promising early evidence of a negative linear correlation between mask use and infection rates to support the claim that masks DO work in mitigating the risk of COVID-19.

Prison Level Relationships

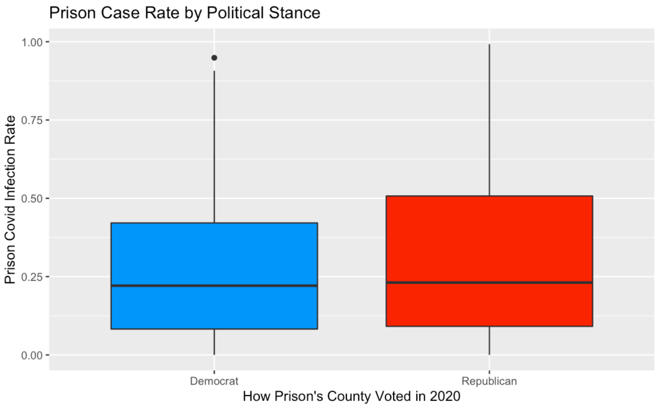
Basic boxplots were also generated to examine the relationship between these variables in prisons. Prisons with large populations seem to have a slightly higher infection rate than prisons with small populations at around .25 and .20 respectively.

Figure 7



Finally, Figure 8 illustrates the relationship between prison infection rates and the political affiliation of their surrounding county. Prisons located in counties that voted Republican in the 2020 election appear to have a slightly higher mean COVID-19 infection rate than the prisons located in counties that voted Democratic. These relationships are explored more in the models.

Figure 8



III. Analysis and Statistical Methods

A. Conditional Probability Models

One initial method used to measure these relationships at the county level was conditional probability models. Conditional probabilities measure the probability that an event occurs, given the probability that another event has already occurred. The conditional probability was utilized to determine if mask usage as well as if the political leaning of a county had an effect on covid case rates.

B. Hypothesis Testing

Hypothesis testing is useful in determining whether the difference between a numerical measure involving two or more groups is statistically significant. In hypothesis testing, a null hypothesis is defined in reference to a particular question that can be answered with statistical analysis. An alternative hypothesis is then declared in opposition to that null hypothesis. The null hypothesis can be rejected only if it falls below a certain significance level. The significance level is the probability that a given result occurs given that the null hypothesis is true. A permutation test is one particular method of hypothesis testing. To conduct a permutation test, there must be two or more categories with corresponding numerical values. The numerical values are then randomly reassigned among the various categories. This random reassignment process is repeated for a large number of iterations to produce a distribution of values. If the original value falls sufficiently within this distribution as determined by the significance level, then the null hypothesis can be accepted.

Within the context of the COVID analysis, there are two different average covid case rates between counties with high mask usage and counties with low mask usage. The null hypothesis is that the mean covid case rate is the same between the two different groups of counties. The alternative hypothesis is that the covid case rate in the group of counties with high mask usage is lower than the group of counties with low mask usage. To test this hypothesis, the actual difference between the average covid case rate of the low mask usage counties and high mask usage counties was calculated. The resulting difference in means was found to be -0.03585878. To conduct the permutations, each covid case rate from a particular county was randomly reassigned to either a "low mask usage" or "high mask usage" county. The difference of the mean of the low mask usage dataset by the high mask usage dataset was then calculated for each permutation. This process was then repeated 1000 times. If the two groups have the same mean case rate, it should follow that the difference between the two means should be 0 on average. The actual difference in the groups' mean case rates must then be evaluated to the distribution of the mean difference in permutations. The exact same process was utilized to compare the average covid death rates between the two groups of counties.

C. Confidence Intervals

A confidence interval defines a range that a given parameter will fall into within a specified probability. In bootstrap sampling, the values from a dataset are resampled with replacement to create a new dataset that is the same size as the original. A statistic from this dataset is then calculated, which is usually the mean or median. The process is then repeated many times, until a sufficient distribution of the calculated statistics is produced.

Within the context of the COVID analysis, confidence intervals were calculated to compare the mean difference in case rates across subsets of the college and prison populations.

Bootstrap sampling was used for this process to find the mean of data subsets, calculate the difference of means, and then generate the 5% and 95% quantiles of that bootstrap mean difference. Bootstrap sampling was used because the college case rate data was not normally distributed and the datasets were large.

Subsets were constructed for the college case rate data for school size, political leaning, and mask usage. The school size subsets consisted of small and large school subsets. The small schools' subset contains the case rates of colleges with less than or equal to 18000 enrolled for 2020. The big schools' subset contains the case rates of colleges with more than 18000 enrolled for 2020. The political leaning subsets consisted of the case rates of colleges in Republican-voting counties and of colleges in Democratic-voting counties in the 2020 presidential election. The mask usage subsets consisted of a high and low mask usage subsets. The high mask usage subset contains case rates of colleges where the surrounding county has a value greater than 0.5 for “Always” wearing a mask. The low mask usage subset contains case rates of colleges where the surrounding county has a value of 0.5 or less for “Always” wearing a mask.

Subsets were constructed of the prison case rate data for prison size, political leaning, and state vs. federal prisons. The prison size subsets consisted of a small and large prison subset. The small prisons subset contains the case rates of prisons with populations of less than 580. The big prisons subset contains the case rates of prisons with populations more than or equal to 580. The political leaning subsets consisted of case rates of prisons in Republican-voting counties and of prisons in Democratic-voting counties. Then there is the state subset and federal subset. The state prisons subset contains the case rates of prisons classified as "State halfway house", "State work camp", "State juvenile detention", "State prison", "State facility", or "State rehabilitation center."

The federal prisons subset contains the case rates of prisons classified as "Detention center", "Juvenile detention at jail", "Federal halfway house", "Low-security work release", "Federal prison", "Reservation jail", or "Jail."

D. Multivariate Regression Analysis

The next model used to analyze these patterns was a multivariate linear regression model, implemented for both college and prison risk factors. Multivariate regression is a key statistical tool to measure how linearly related a predictor variable is with a response variable. When examining risk factors for COVID-19 case rates in U.S. colleges and prisons, it is a powerful way to examine how cases change based on different predictor variables such as population size, political leaning, and mask usage.

For COVID-19 cases in U.S. colleges and universities, the core model used was

$$\begin{aligned} Case\_rate_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 * State_i + \boldsymbol{\beta}_2 * Rank2020_i + \boldsymbol{\beta}_3 * Enrollment2020_i \\ + \boldsymbol{\beta}_4 * Mask\ Usage_i + \boldsymbol{\beta}_5 * Election\ Winner_i \end{aligned} \tag{1}$$

where, for each college, the number of COVID-19 cases per 1,000 people was regressed against the college's state, its 2020 US News Ranking, its 2020 enrollment, mask usage of the surrounding county, and the political affiliation of the surrounding county based on 2020 election results. The state vector was a series of dummy variables for each state where "Alabama" was its reference category. Mask usage was the percentage of the county that "never wears masks" while rank and enrollment were continuous variables. Election winner was a categorical variable as well - taking on "Republican" or "Democrat" as its levels - and thus Democrat was the reference category.

A multivariate linear regression model was again used to dive deeper into risk factors of U.S. prisons and penitentiary COVID-19 exposure. The main model generated was

$$\begin{aligned} Case\_rate_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 * Facility\ Type_i + \boldsymbol{\beta}_2 * State_i + \boldsymbol{\beta}_3 * Population_i \\ + \ \boldsymbol{\beta}_4 * Political\ Party_i + \boldsymbol{\beta}_5 * County\ Population_i \end{aligned} \tag{2}$$

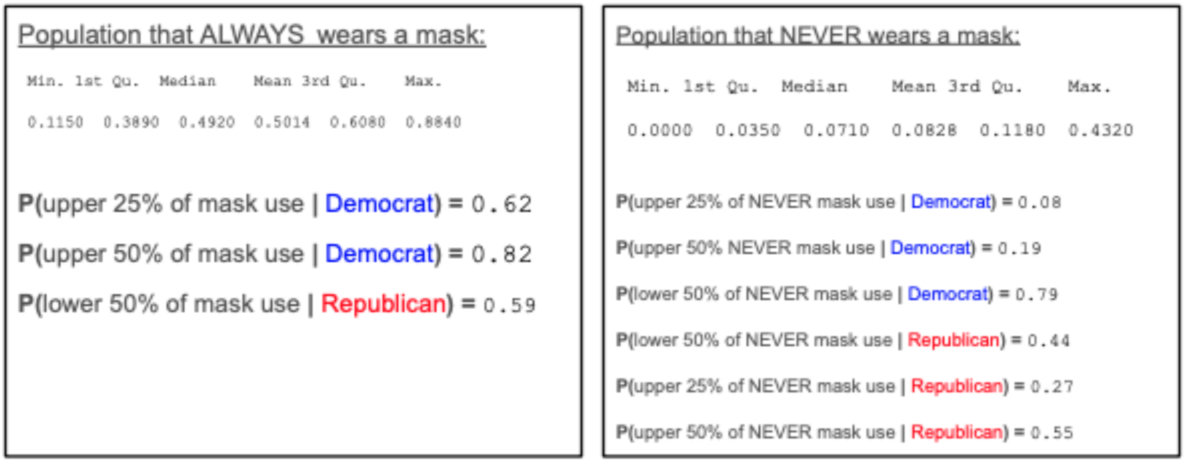
where case rate was the number of COVID-19 cases per 1,000 people, facility type included an array of various state operated, federally operated, high security, and low-security prisons, and state was a vector of all US states using Alabama as the reference category. Population refers to the latest inmate population at each prison, political party refers to the political leanings of the surrounding county based on 2020 election results, and county population is defined as the population of the surrounding county.

IV. Results

A. Conditional Probability Models

The results of the calculated conditional probabilities are shown below. These boxes display conditional probabilities based on the percentage of a county's population that reported that they either always wear a mask in public or that they never wear a mask in public. The two percentages, “never” and “always,” are two separate variables, and the distribution of each is given in their corresponding box. The box on the left shows data for the county’s population that always wears a mask in public. By comparing the distribution of populations that always wear masks to how their county voted in the 2020 presidential election, we were able to calculate the conditional probability that a county had a high instance of mask usage given they voted either Republican or Democrat. For instance, the probability that a county is in the top 50% that reports always wearing masks given that they voted Democrat is 82%. Conditional probabilities were also calculated for counties based on the amount of their population that reported never wearing masks. By looking at the conditional probabilities of how much a county's population reports either always wearing masks or never wearing masks, given they voted Republican or Democrat, we see that counties voted Democrat are more likely to wear masks, whereas counties with low mask usage are more likely to have voted Republican.

Figure 9



Population that ALWAYS wears a mask:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1150  0.3890  0.4920  0.5014  0.6080  0.8840
```

P(upper 25% of mask use | Democrat) = 0.62

P(upper 50% of mask use | Democrat) = 0.82

P(lower 50% of mask use | Republican) = 0.59

Population that NEVER wears a mask:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0350  0.0710  0.0828  0.1180  0.4320
```

P(upper 25% of NEVER mask use | Democrat) = 0.08

P(upper 50% NEVER mask use | Democrat) = 0.19

P(lower 50% of NEVER mask use | Democrat) = 0.79

P(lower 50% of NEVER mask use | Republican) = 0.44

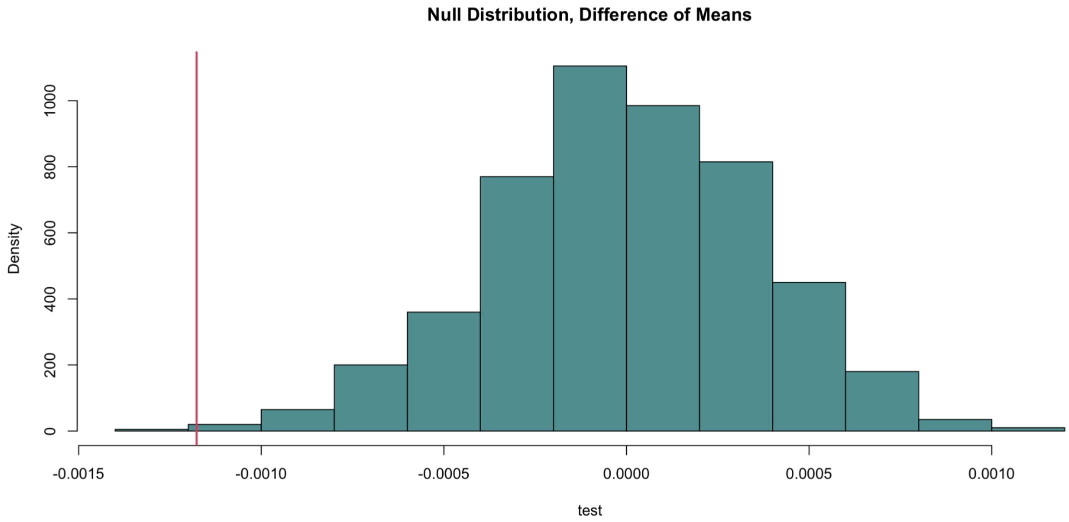
P(upper 25% of NEVER mask use | Republican) = 0.27

P(upper 50% of NEVER mask use | Republican) = 0.55

B. Hypothesis Testing

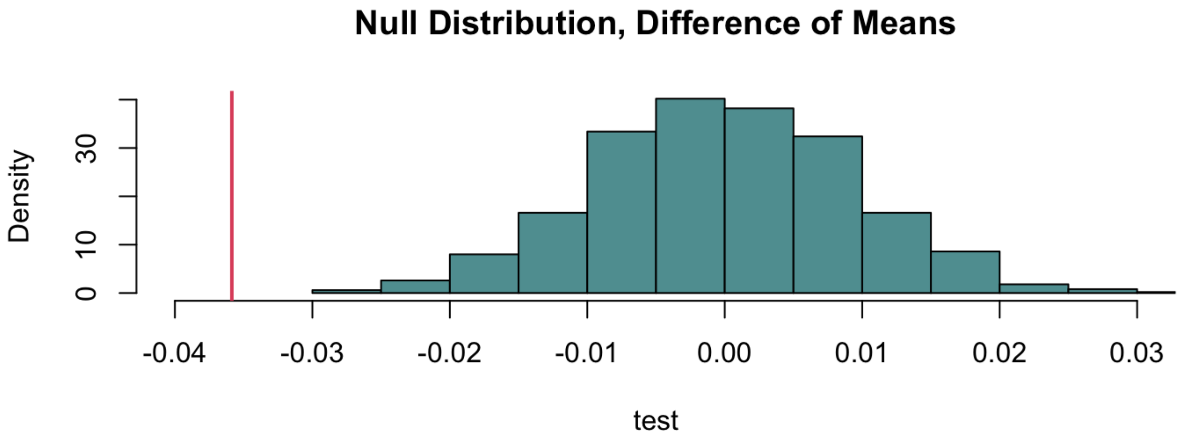
The distribution of the difference of means from the random permutations is displayed by the histogram below. The actual difference in means between the mean case rate of the low mask usage group and the high mask usage group is marked by the red line. The result is that the actual mean result was larger than 0.1% of the means from the random permutations. This is less than the significance level of 5%. Thus, the null hypothesis can be rejected and the alternative can be accepted, and therefore the conclusion can be reached that the mean case rate in counties with low mask usage was lower than that of counties with high mask usage.

Figure 10



The distribution of the result of the same process for COVID deaths between the two groups of counties is displayed below. Zero of the mean differences from the random permutations were greater than the actual difference of the counties. This leaves us with a p-value of 0, which is lower than the significance level of 5%. Since the p-value is low, we can reject the null hypothesis and accept the alternative, and conclude that the mean covid death rate is lower in counties with higher mask use.

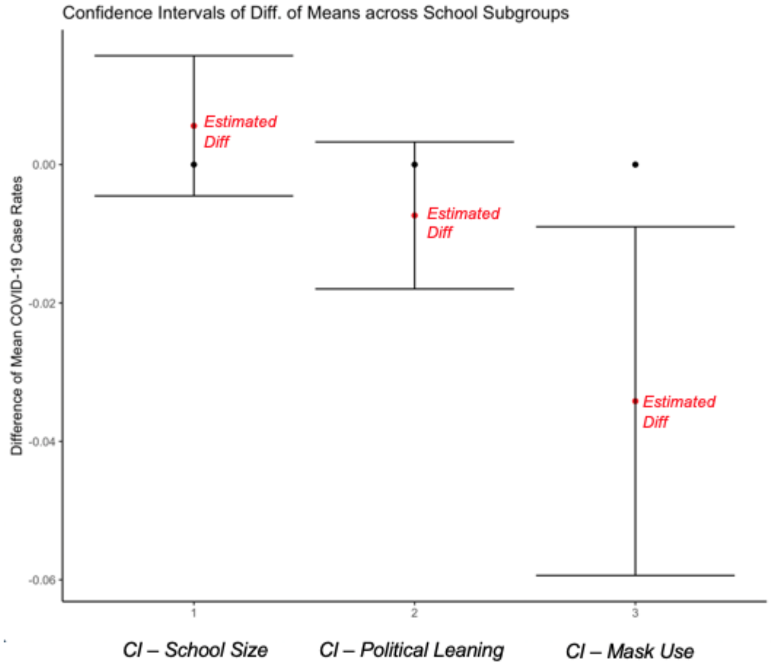
Figure 11



C. Confidence Intervals

The 95% confidence interval for the mean difference in college case rates between big and small schools is ***(-0.00453, 0.01572).*** Zero is contained in this confidence interval. This suggests that there isn't a difference between the mean of case rates from large colleges and the mean of case rates from small colleges. There is 95% certainty that this confidence interval contains the true difference in mean of case rates between the large and small schools.

Figure 12



Confidence Intervals Plot: The lines represent the confidence intervals for each variable subset. The black points show where zero is in relation to the confidence interval.

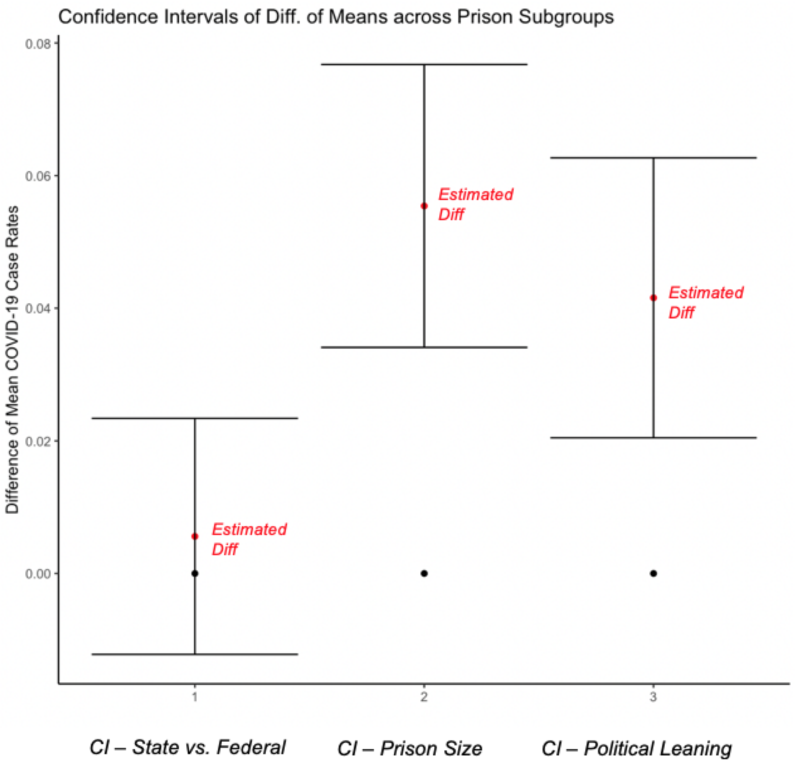
The 95% confidence interval for the mean difference in college case rates between colleges in Republican-voting counties and of colleges in Democratic-voting counties is ***(-0.01797, 0.00326).*** Zero is contained in this confidence interval. This suggests that there isn't a difference between the mean of college case rates of colleges in counties that voted Democrat in 2020 and the mean of the college case rates of colleges in counties that voted Republican in 2020.

The 95% confidence interval for the mean difference in college case rates between colleges in high mask usage counties and of colleges in low mask usage counties is ***(-0.04140, -0.00899).*** Zero is *not* contained in this confidence interval. This implies that there is a difference between the mean of case rates of colleges in high mask usage counties and the mean of case rates of colleges in low mask usage counties. This suggests that colleges in counties with low mask usage have higher case rates than the colleges in counties with high mask usage.

The 95% confidence interval for the mean difference in prison case rates between big and small prisons is ***(0.03409, 0.07675).*** Zero is *not* contained in this confidence interval. This implies that there is a difference between the mean of case rates from large prisons and the mean of case rates from small prisons. This suggests that large prisons have higher case rates than small prisons.

The 95% confidence interval for the mean difference in prison case rates between prisons in Republican-voting counties and of prisons in Democratic-voting counties is ***(0.02046, 0.06268).*** Zero is *not* contained in this confidence interval. This implies that there is a difference between the mean of prison case rates of prisons in counties that voted Democrat in 2020 and the mean of the prison case rates of prisons in counties that voted Republican in 2020. This suggests that prisons in counties that voted Republican have higher case rates than prisons in counties that voted Democratic.

Figure 13



Confidence Intervals Plot: The lines represent the confidence intervals for each variable subset. The black points show where zero is in relation to the confidence interval.

The 95% confidence interval for the mean difference in prison case rates between state and federal prisons is ***(-0.08211, 0.04652).*** Zero is contained in this confidence interval. This suggests that there isn't a difference between the mean of case rates in prisons previously classified as state prisons and the mean of case rates in prisons previously classified as federal prisons.

These confidence intervals suggest that mask usage, prison size and the political leaning of the county the prison is located in are significant factors for the vulnerability of covid infection rate. Students at colleges in counties with low mask usage are at a greater risk for covid infection than students at colleges in counties with high mask usage. Inmates at big prisons are more at risk of catching covid than at small prisons. Prisons in counties that voted republican in the 2020 election are at risk for higher inmate covid infections than prisons in democratic voting counties. Mask usage, population size, and political leaning, are significant vulnerability factors that should be taken note of.

D. Multivariate Regression Analysis

Results from regression equation (1) are shown to the left. A sampling of key predictor variables are included based on their statistical significance and economic predictive power and the full list of variables with their coefficients can be found in the R Markdown file.

College location (by state) was important in predicting its case rate. Having a college in the state of California was associated with 87.4 fewer cases/1,000 people than the reference category of Alabama while a college in the state of South Carolina was associated with 220.4 more cases/1,000 people than Alabama, holding all other variables fixed. 16 of 50 states were statistically significant predictors of case rate, including California, Colorado, Florida, Hawaii, Illinois, Kansas, Nevada, New Hampshire, New Jersey, New York, North Dakota, Oregon, South Carolina, Texas, Washington, and Wyoming - all of which have lower case rates than Alabama except for North Dakota, South Carolina, and Wyoming (which have higher case rates). The 2020 rank was not a statistically significant predictor of a college's case rate nor was the political leaning of the county as well as the college's 2020 enrollment. Additionally, holding all other variables constant, for each percentage point increase in the number of people in a county that never wore a mask, there is a statistically significant correlation to a 417.5 decrease in cases/1,000 people.

Regression Table 1 – College COVID-19 Data
(SE in parentheses)
Dependent Variable = Rate of COVID-19 Cases/1,000 people
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

	(1)
California	-87.4*** (24.85)
South Carolina	220.4*** (45.09)
Florida	-57.92* (23.52)
Illinois	-56.71* (28.34)
New Jersey	-58.30* (28.49)
New York	-61.76* (28.49)
2020 Enrollment	$3.204\text{x}10^{-5}$ ($5.06\text{x}10^{-4}$)
2020 Rank	-0.02937 (0.04145)
% Never Wears Masks	-417.5* (169.01)
2020 Election Winner	-0.3669 (9.959)
Intercept (B_0)	128.4** (47.46)
R^2 - overall	.4568
R^2 - adjusted	.2942
F-Statistic	2.809***

Thus, key results are that states had a high correlation to the COVID-19 case rate on college campuses and universities, holding other variables constant. Though population was found to be highly correlated with COVID-19 cases in other models, enrollment was not here - likely due to the normalization of case rate (instead of looking at total cases). The lack of influence of political leanings is likely attributed to the fact that political data was taken at the county level while most universities enacted different COVID-19 policies than their local government. Colleges, especially private colleges, were able to freely make rules and mitigation plans that differed from their local town or city and generally have insular campuses with limited contact/transmission spread to the outside community.

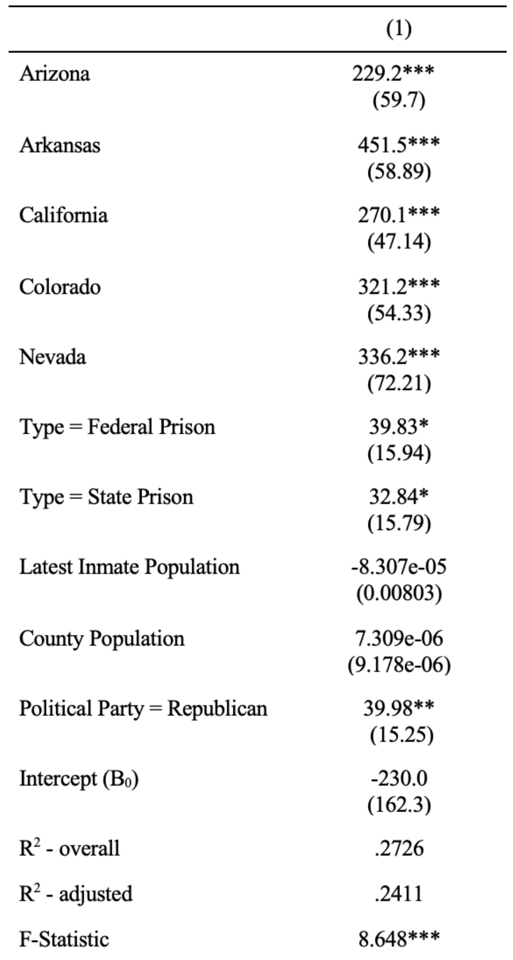
The adjusted R^2 value is .2942, indicating that 29.42% of a college's case rate can be explained by these attributes. This is fairly large for just a few variables so a college's location, enrollment, rank, local political leaning, and local mask use account for 29.42% of the variation in case rate. This is supported by an F statistic of 2.809 on 50 and 167 DF and a p-value of 4.27×10^{-7} - allowing us to reject the null hypothesis that the joint predictive power of these variables is 0.

Prison location (by state) was again important in predicting its case rate. Holding other variables constant, having a prison in the state of California was associated with 270.1 more cases/1,000 people than the reference category of Alabama (statistically significant at the 1% level) while a prison in the state of Nevada was associated with 336.2 more cases/1,000 people than Alabama (significant at the 1% level). 40 of 50 states were statistically significant predictors of case rate, indicating that location of a prison has a strong correlation to COVID-19 case rates. Latest inmate population was not a statistically significant predictor of a prison's case rate nor was the population of the surrounding count. However, political party *was*, as prisons in

Republican counties were correlated with 39.98 more cases/1,000 people than in Democratic counties (again holding other variables fixed).

Thus, key results were that COVID-19 cases in prisons were correlated with their given state and the surrounding politics of their county. Inmate population and county population were not statistically significantly correlated with COVID-19 cases; which is at least partly due to the examination of *case rate* rather than total cases. However, it was still hypothesized that larger facilities would have higher rates of transmission which is fairly interesting to see rejected by this specific model.

Political party was a strong predictor of a prison's case rate, with Republican counties seeing higher COVID-19 cases in their prisons. This may be due to specific legislation being passed in these counties in regards to vaccinations, COVID-19 testing and quarantine protocols, and healthcare access provided to each prison by its governing body.



Regression Table 2 – Prison COVID-19 Data Key Coefficients
(SE in parentheses)
Dependent Variable = Rate of COVID-19 Cases/1,000 people
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

	(1)
Arizona	229.2*** (59.7)
Arkansas	451.5*** (58.89)
California	270.1*** (47.14)
Colorado	321.2*** (54.33)
Nevada	336.2*** (72.21)
Type = Federal Prison	39.83* (15.94)
Type = State Prison	32.84* (15.79)
Latest Inmate Population	-8.307e-05 (0.00803)
County Population	7.309e-06 (9.178e-06)
Political Party = Republican	39.98** (15.25)
Intercept (B_0)	-230.0 (162.3)
R^2 - overall	.2726
R^2 - adjusted	.2411
F-Statistic	8.648***

The adjusted R^2 value is .2411, indicating that 24.11% of a prison's case rate can be explained by these attributes. Thus, a prison's location and local politics account for 24.11% of the variation in case rate - which is surprisingly high. This is supported by an F statistic of 8.648 on 62 and 1,431 DF and a p-value of $2.2\text{x}10^{-16}$ - allowing us to reject the null hypothesis that the joint predictive power of these variables is 0.

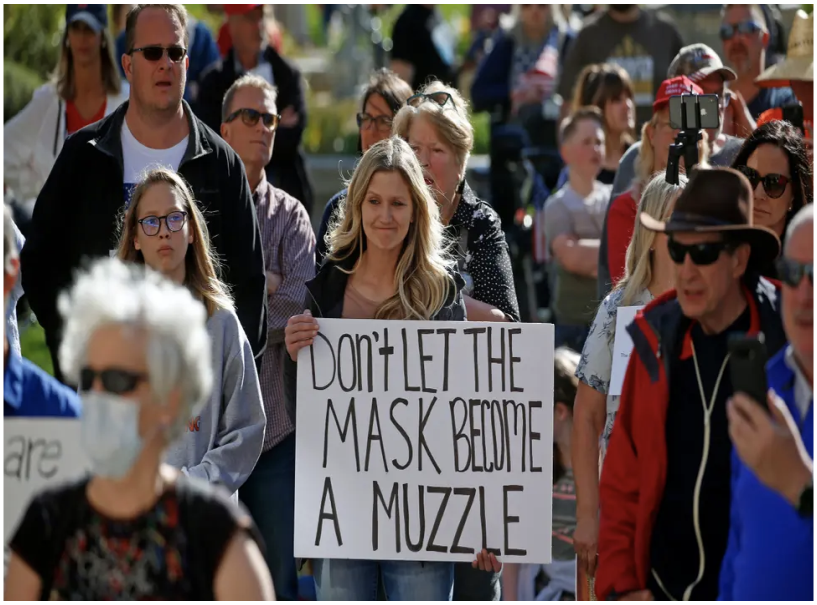
Conclusions

COVID-19 has spread like wildfire throughout America and will be remembered as a once in a generation global event. As Americans grapple with the economic and health impacts of the virus and attempt to return to normalcy the question of who is most vulnerable to infection remains highly important. Determining which subsets of the population experience higher than average infection rates and why their risk is so much higher allows for implementation of mitigation protocols such as mask mandates if and where they are effective.

The impact of population size on case rates differed in the analysis and thus there is inconclusive evidence on its impact. For example, in counties and schools there is no clear evidence that larger populations experience worse rates of COVID-19 whereas in prisons large facilities did see higher rates of infection than small facilities. Prisons, given their close quarters and notoriously large incarcerated populations, were deemed overcrowded “epicenters” of spread by a recent Northwestern and World Bank study (NPR 2021). This same study also suggests that a reduction in the U.S. jail population could have prevented “tens of thousands of deaths” and millions of cases (NPR 2021). Colleges, on the other hand, were likely incentivized to prevent as many cases as possible given that their students were paying clientele and they thus implemented strict testing and quarantine protocols to counteract the transmission that happens with large student populations.

The effectiveness of mask usage was made clear via all models used. On average, the case rate in counties, colleges, and prisons where mask usage was high was lower than in locations where masks were used less frequently. This points to the high degree of political polarization the country has experienced around COVID-19 and mask use specifically. Many parts of the country, largely concentrated in Republican leaning states, have taken issue with mask wearing and mask mandates due to a wave of cynicism towards science and technology. This distrust of leading experts largely comes from messaging put forth by President Donald Trump when he was in office in 2020 as well as ongoing anti-vaxxer sentiment that has been growing in the country for decades. Shown below is an image from 2020 where many gathered to protest the use of masks in the fight against the pandemic. It's become a cultural point of division that - given the analysis and results here - puts anti- maskers at a heightened risk of infection.

Figure 14


Don't LET THE MASK BECOME A MUZZLE
are

It is clear that people in America are facing different risk profiles based on their location. COVID-19, as it continues to mutate and evade vaccination efforts, will remain a public threat into the foreseeable future. As society and individuals continue to grapple with the risks, governments, colleges, and prisons can use the analysis performed here to better understand the risk factors associated with infection and work to prevent as many unnecessary infections and deaths as possible.

Sources

1. Bundgaard, Henning, Johan Skov Bundgaard, Daniel Emil Tadeusz Raaschou-Pedersen, Christian von Buchwald, Tobias Todsen, Jakob Boesgaard Norsk, Mia M. Pries-Heje et al. "Effectiveness of adding a mask recommendation to other public health measures to prevent SARS-CoV-2 infection in Danish mask wearers: a randomized controlled trial." *Annals of Internal Medicine* 174, no. 3 (2021): 335-343.
2. Chappell, Bill. "Crowded U.S. Jails Drove Millions Of COVID-19 Cases, A New Study Says." NPR. September 02, 2021. Accessed December 08, 2021. https://www.npr.org/2021/09/02/1033326204/crowded-jails-drove-millions-of-covid-19-cases-a-new-study-says.
3. “Coronavirus World Map: Tracking the Global Outbreak.” *The New York Times*, The New York Times, 28 Jan. 2020, https://www.nytimes.com/interactive/2021/world/covid-cases.html.
4. Elias, Thomas D. "Thomas D. Elias: Anti-maskers, Anti-vaxxers: What's the Difference?" Napa Valley Register. October 27, 2020. Accessed December 08, 2021. https://napavalleyregister.com/opinion/columnists/thomas-d-elias-anti-maskers-anti-vaxxers-what-s-the-difference/article_c169db16-c05d-5564-9913-51a6f2b7e48d.html.
5. Ettner, Susan L. "New evidence on the relationship between income and health." *Journal of health economics* 15, no. 1 (1996): 67-85.
6. Mayo Clinic Staff. "COVID-19 (coronavirus): Long-term Effects." Mayo Clinic. October 22, 2021. Accessed December 05, 2021. https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-long-term-effects/art-20490351.
7. MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V9, UNF:6:qSwUYo7FKxI6vd/3Xev2Ng== [fileUNF]
8. "The New York Times. (2021). Coronavirus (Covid-19) Data in the United States, from https://github.com/nytimes/covid-19-data."
9. StateUniversity. "College Rankings." StateUniversity.com. Accessed December 08, 2021. https://www.stateuniversity.com/rank/tot_enroll_rank/25.

Appendix

Link to RMD output:
https://github.com/austineaton/GU-ANLY511-FinalProject/blob/main/Final_Report_Code/Final_Report_Code.pdf

Final Report

Christopher Delaney, Austin Eaton, Clara Richter, Elise Rust

12/8/2021

All files used in this analysis are available in the following GitHub repository: https://github.com/austineaton/GU-ANLY511-FinalProject

Libraries Used in Analysis

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readr)  # for read_csv
library(knitr)  # for kable
library(usmap)
library(formatR)
```

1. Data Cleaning

a. Colleges

```
colleges <- read.csv("../colleges/colleges.csv")
attendance <- read.csv("../college_attendance.csv", na.strings=c("","NA"))
mask_use <- read.csv("../Final_Report_Code/MaskUse.csv")
election_results <- read.csv("../countypres_2000-2020.csv")
```

```
head(colleges)
```

```
##         date   state     county       city ipeds_id
## 1 2021-05-26 Alabama    Madison Huntsville   100654
## 2 2021-05-26 Alabama Montgomery Montgomery   100724
## 3 2021-05-26 Alabama  Limestone     Athens   100812
## 4 2021-05-26 Alabama        Lee     Auburn   100858
## 5 2021-05-26 Alabama Montgomery Montgomery   100830
## 6 2021-05-26 Alabama     Walker     Jasper   102429
##                         college cases cases_2021 notes
## 1          Alabama A&M University    41         NA
## 2        Alabama State University     2         NA
## 3         Athens State University    45         10
## 4               Auburn University  2742        567
## 5 Auburn University at Montgomery   220         80
## 6  Bevill State Community College     4         NA
```

```
head(attendance)
```

```
##   Rank_2020 Rank_2015                               Name Enroll_2020
## 1         1         1 University of Phoenix-Online Campus     212,044
## 2         2        25        Western Governors University     121,437
## 3         3       111   Southern New Hampshire University     104,068
## 4         4       N/A       University of Phoenix-Arizona      95,777
## 5         5        11              Grand Canyon University      90,253
## 6         6         3                  Liberty University      79,152
##     Enroll_2015
## 1     No Change
## 2        46,733
## 3        28,389
## 4 Not Available
## 5        55,497
## 6        77,338
```

```
head(mask_use)
```

```
##   fips NEVER RARELY SOMETIMES FREQUENTLY ALWAYS   state abbr  county
## 1 1001 0.053  0.074     0.134      0.295  0.444 Alabama   AL Autauga
## 2 1003 0.083  0.059     0.098      0.323  0.436 Alabama   AL Baldwin
## 3 1005 0.067  0.121     0.120      0.201  0.491 Alabama   AL Barbour
## 4 1007 0.020  0.034     0.096      0.278  0.572 Alabama   AL    Bibb
## 5 1009 0.053  0.114     0.180      0.194  0.459 Alabama   AL  Blount
## 6 1011 0.031  0.040     0.144      0.286  0.500 Alabama   AL Bullock
```

```
head(election_results)
```

```
##   year   state state_po county_name county_fips    office      candidate
## 1 2000 ALABAMA       AL     AUTAUGA        1001 PRESIDENT        AL GORE
## 2 2000 ALABAMA       AL     AUTAUGA        1001 PRESIDENT GEORGE W. BUSH
## 3 2000 ALABAMA       AL     AUTAUGA        1001 PRESIDENT    RALPH NADER
## 4 2000 ALABAMA       AL     AUTAUGA        1001 PRESIDENT          OTHER
## 5 2000 ALABAMA       AL     BALDWIN        1003 PRESIDENT        AL GORE
## 6 2000 ALABAMA       AL     BALDWIN        1003 PRESIDENT GEORGE W. BUSH
##        party candidatevotes totalvotes  version  mode
## 1   DEMOCRAT           4942      17208 20191203 TOTAL
## 2 REPUBLICAN          11993      17208 20191203 TOTAL
## 3      GREEN            160      17208 20191203 TOTAL
## 4      OTHER            113      17208 20191203 TOTAL
## 5   DEMOCRAT          13997      56480 20191203 TOTAL
## 6 REPUBLICAN          40872      56480 20191203 TOTAL
```

```
# Basic summary statistics and structure
summary(colleges)
```

```
##      date              state              county              city          
##  Length:1948        Length:1948        Length:1948        Length:1948       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##    ipeds_id           college              cases           cases_2021    
##  Length:1948        Length:1948        Min.   :   0.0   Min.   :   0.0  
##  Class :character   Class :character   1st Qu.:  32.0   1st Qu.:  23.0  
##  Mode  :character   Mode  :character   Median : 114.5   Median :  65.0  
##                                        Mean   : 363.5   Mean   : 168.1  
##                                        3rd Qu.: 303.0   3rd Qu.: 159.0  
##                                        Max.   :9914.0   Max.   :3158.0  
##                                                         NA's   :337     
##     notes          
##  Length:1948       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
## 
```

```
summary(attendance)
```

```
##   Rank_2020          Rank_2015             Name            Enroll_2020       
##  Length:500         Length:500         Length:500         Length:500        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##  Enroll_2015       
##  Length:500        
##  Class :character  
##  Mode  :character  
```

```
summary(mask_use)
```

```
##       fips           NEVER             RARELY          SOMETIMES     
##  Min.   : 1001   Min.   :0.00000   Min.   :0.00000   Min.   :0.0010  
##  1st Qu.:18178   1st Qu.:0.03400   1st Qu.:0.04000   1st Qu.:0.0790  
##  Median :29176   Median :0.06800   Median :0.07300   Median :0.1150  
##  Mean   :30384   Mean   :0.07994   Mean   :0.08292   Mean   :0.1213  
##  3rd Qu.:45080   3rd Qu.:0.11300   3rd Qu.:0.11500   3rd Qu.:0.1560  
##  Max.   :56045   Max.   :0.43200   Max.   :0.38400   Max.   :0.4220  
##    FREQUENTLY         ALWAYS          state               abbr          
##  Min.   :0.0290   Min.   :0.1150   Length:3142        Length:3142       
##  1st Qu.:0.1640   1st Qu.:0.3932   Class :character   Class :character  
##  Median :0.2040   Median :0.4970   Mode  :character   Mode  :character  
##  Mean   :0.2077   Mean   :0.5081                                        
##  3rd Qu.:0.2470   3rd Qu.:0.6138                                        
##  Max.   :0.5490   Max.   :0.8890                                        
##     county         
##  Length:3142       
##  Class :character  
##  Mode  :character  
##                    
##                    
## 
```

```
summary(election_results)
```

```
##       year         state             state_po         county_name       
##  Min.   :2000   Length:72617       Length:72617       Length:72617      
##  1st Qu.:2004   Class :character   Class :character   Class :character  
##  Median :2012   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :2011                                                           
##  3rd Qu.:2020                                                           
##  Max.   :2020                                                           
##                                                                         
##   county_fips       office           candidate            party          
##  Min.   : 1001   Length:72617       Length:72617       Length:72617      
##  1st Qu.:18103   Class :character   Class :character   Class :character  
##  Median :29205   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :30264                                                           
##  3rd Qu.:45057                                                           
##  Max.   :56045                                                           
##  NA's   :57                                                              
##  candidatevotes      totalvotes         version              mode          
##  Min.   :      0   Min.   :      0   Min.   :20191203   Length:72617      
##  1st Qu.:    119   1st Qu.:   5177   1st Qu.:20191203   Class :character  
##  Median :   1305   Median :  11194   Median :20191203   Mode  :character  
##  Mean   :  10843   Mean   :  42510   Mean   :20197111                     
##  3rd Qu.:   5898   3rd Qu.:  29874   3rd Qu.:20210622                     
##  Max.   :3028885   Max.   :4264365   Max.   :20210622                     
##  NA's   :405       NA's   :5                                            
```

```
## # 1. Fix column names
names(attendance) <- c("rank2020", "rank2015", "college", "enroll2020", "enroll2015")
names(election_results) <- c("year", "state", "state_abbr", "county", "fips",
                              "office", "candidate", "party", "candidatevotes",
                              "totalvotes", "version", "mode")

## 2. Remove unnecessary columns
# colleges --> remove cases2021 (only interested in total cases) and
#              notes (extraneous), date (all on 5/26/2021)
colleges <- colleges %>%
  select(-c(cases_2021, notes, date))

# mask_use --> remove fips code (not used for join), abr (not useful)
mask_use <- mask_use %>%
  select(-c(fips, abbr))

# election_results --> remove state_abbr, fips, office, totalvotes, version, mode
election_results <- election_results %>%
  filter(year == 2020) %>%  # just 2020 results
  select(-c(year, state_abbr, fips, office, totalvotes, version, mode))


## 4. Missing/incorrect values
# Colleges dataframe
sum(is.na(colleges$cases) == TRUE)
```

```
## [1] 0
```

```
sum(is.na(colleges$college) == TRUE)
```

```
## [1] 0
```

```
colleges$ipeds_id[colleges$ipeds_id=="laccd"] <- NA
# No NULL values in colleges dataframe

# Attendance dataframe
sum(is.na(attendance)) # 31 NAs
```

```
## [1] 2
```

```
attendance$rank2015[attendance$rank2015=="N/A"] <- NA
attendance$enroll2015[attendance$enroll2015=="Not Available"] <- NA
attendance$enroll2015[attendance$enroll2015=="No Change"] <- NA

sum(is.na(attendance)) # 20 NAs
```

```
## [1] 20
```

```
# Mask Use dataframe
sum(is.na(mask_use)) # No NAs
```

```
## [1] 0
```

```
# ElectionResults dataframe
sum(is.na(election_results)) # 1 NA
```

```
## [1] 1
```

```
election_results <- election_results %>%
  drop_na(candidatevotes)

# FIX NAs
sum(is.na(attendance$enroll2020))
```

```
## [1] 1
```

```
# Southeastern Louisiana had 2020 enrollment of 13,490
attendance$enroll2020[attendance$college == "Southeastern Louisiana University"] = 13490
# Drop entirely blank rows at bottom
attendance <- attendance %>%
  drop_na(rank2020)

# Remove extra whitespace so type conversion can happen
attendance$enroll2020 = str_trim(attendance$enroll2020, side = c("both"))
attendance$enroll2015 = str_trim(attendance$enroll2015, side = c("both"))
attendance$rank2020 = str_trim(attendance$rank2020, side = c("both"))
attendance$rank2015 = str_trim(attendance$rank2015, side = c("both"))

# Remove commas from numbers
attendance$enroll2020 <- as.numeric(gsub(",","",attendance$enroll2020))
attendance$enroll2015 <- as.numeric(gsub(",","", attendance$enroll2015))

## 3. Correct datatypes before median imputation
str(colleges)
```

```
## 'data.frame':    1948 obs. of  6 variables:
##  $ state   : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ county  : chr  "Madison" "Montgomery" "Limestone" "Lee" ...
##  $ city    : chr  "Huntsville" "Montgomery" "Athens" "Auburn" ...
##  $ ipeds_id: chr  "100654" "100724" "100812" "100858" ...
##  $ college : chr  "Alabama A&M University" "Alabama State University" "Athens State University" "Auburn Univer
sity" ...
##  $ cases   : int  41 2 45 2742 220 4 263 137 49 76 ...
```

```
str(attendance)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ rank2020  : chr  "1" "2" "3" "4" ...
##  $ rank2015  : chr  "1" "25" "111" NA ...
##  $ college   : chr  "University of Phoenix-Online Campus" "Western Governors University" "Southern New Hampshi
re University" "University of Phoenix-Arizona" ...
##  $ enroll2020: num  212044 121437 104068 95777 90253 ...
##  $ enroll2015: num  NA 46733 28389 NA 55497 ...
```

```
colleges <- colleges %>%
  mutate(ipeds_id = as.numeric(colleges$ipeds_id)) %>%
  mutate(cases = as.numeric(cases))

attendance <- attendance %>%
  mutate(rank2020 = as.numeric(rank2020)) %>%
  mutate(rank2015 = as.numeric(rank2015))


# Median imputation for enroll_2015
attendance$enroll2015[is.na(attendance$enroll2015)] <- median(attendance$enroll2015, na.rm = TRUE)

sum(is.na(attendance$enroll2015))
```

```
## [1] 0
```

```
###### Clean Election Results further to prep for join
# Fix capitalization of election results
election_results <- election_results %>%
  mutate(state = str_to_title(state)) %>%
  mutate(county = str_to_title(county)) %>%
  mutate(candidate = str_to_title(candidate)) %>%
  mutate(party = str_to_title(party))

head(election_results)
```

```
##     state  county         candidate      party candidatevotes
## 1 Alabama Autauga Joseph R Biden Jr   Democrat           7503
## 2 Alabama Autauga             Other      Other            429
## 3 Alabama Autauga    Donald J Trump Republican          19838
## 4 Alabama Baldwin Joseph R Biden Jr   Democrat          24578
## 5 Alabama Baldwin             Other      Other           1557
## 6 Alabama Baldwin    Donald J Trump Republican          83544
```

```
# Calculate one winner for each county
election_results <- election_results %>%
  group_by(county,state) %>%
  mutate(winner = party[which.max(candidatevotes)])

# Keep only row of winner for each county
election_results <- election_results %>%
  group_by(county, state) %>%
  filter(party == winner) %>%
  select(-c(candidate, party, candidatevotes))

head(election_results)
```

```
## # A tibble: 6 × 3
## # Groups:   county, state [6]
##   state   county  winner    
##   <chr>   <chr>   <chr>     
## 1 Alabama Autauga Republican
## 2 Alabama Baldwin Republican
## 3 Alabama Barbour Republican
## 4 Alabama Bibb    Republican
## 5 Alabama Blount  Republican
## 6 Alabama Bullock Democrat
```

```
##################
# Join
colleges_full <- inner_join(colleges, attendance, by = "college" )
colleges_full <- left_join(colleges_full, mask_use, by = c("county", "state"))
colleges_full <- left_join(colleges_full, election_results, by = c("county", "state"))
head(colleges_full)
```

```
##     state                 county       city ipeds_id
## 1 Alabama                    Lee     Auburn   100858
## 2 Alabama                   Pike       Troy   102368
## 3 Alabama              Jefferson Birmingham   100663
## 4 Alabama                 Mobile     Mobile   102094
## 5  Alaska Anchorage Municipality  Anchorage   102553
## 6 Arizona               Maricopa    Phoenix   104717
##                                college cases rank2020 rank2015 enroll2020
## 1                    Auburn University  2742      101      150      30440
## 2                      Troy University   855      285      217      16981
## 3 University of Alabama at Birmingham  3143      193      256      21923
## 4         University of South Alabama   780      338      346      14834
## 5      University of Alaska Anchorage   186      348      283      14644
## 6              Grand Canyon University  1935        5       11      90253
##   enroll2015 NEVER RARELY SOMETIMES FREQUENTLY ALWAYS     winner
## 1      24864 0.053  0.064     0.138      0.183  0.562 Republican
## 2      20573 0.060  0.131     0.145      0.333  0.330 Republican
## 3      18568 0.049  0.037     0.107      0.179  0.628   Democrat
## 4      15065 0.077  0.070     0.102      0.244  0.506 Republican
## 5      17363 0.042  0.050     0.049      0.196  0.663       <NA>
## 6      55497 0.023  0.025     0.059      0.158  0.734   Democrat
```

```
str(colleges_full)
```

```
## 'data.frame':    354 obs. of  16 variables:
##  $ state     : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ county    : chr  "Lee" "Pike" "Jefferson" "Mobile" ...
##  $ city      : chr  "Auburn" "Troy" "Birmingham" "Mobile" ...
##  $ ipeds_id  : num  100858 102368 100663 102094 102553 ...
##  $ college   : chr  "Auburn University" "Troy University" "University of Alabama at Birmingham" "University of
South Alabama" ...
##  $ cases     : num  2742 855 3143 780 186 ...
##  $ rank2020  : num  101 285 193 338 348 5 5 5 94 94 ...
##  $ rank2015  : num  150 217 256 346 283 11 11 11 135 135 ...
##  $ enroll2020: num  30440 16981 21923 14834 14644 ...
##  $ enroll2015: num  24864 20573 18568 15065 17363 ...
##  $ NEVER     : num  0.053 0.06 0.049 0.077 0.042 0.023 0.023 0.023 0.039 0.039 ...
##  $ RARELY    : num  0.064 0.131 0.037 0.07 0.05 0.025 0.025 0.025 0.08 0.08 ...
##  $ SOMETIMES : num  0.138 0.145 0.107 0.102 0.049 0.059 0.059 0.059 0.07 0.07 ...
##  $ FREQUENTLY: num  0.183 0.333 0.179 0.244 0.196 0.158 0.158 0.158 0.11 0.11 ...
##  $ ALWAYS    : num  0.562 0.33 0.628 0.506 0.663 0.734 0.734 0.734 0.702 0.702 ...
##  $ winner    : chr  "Republican" "Republican" "Democrat" "Republican" ...
```

```
## Calculate "rate" parameter and percent change in enrollment
colleges_full <- colleges_full %>%
  mutate(perc_enroll_change = (enroll2020 - enroll2015)/enroll2015) %>%
  mutate(case_rate = cases/enroll2020)

# Remove any duplicate rows
colleges_full = colleges_full[!duplicated(colleges_full$ipeds_id), ]


# write.csv(colleges_full, "~/Desktop/colleges_cleaned2.csv")
```

b. Mask Use

```
# read in county mask use data
myfile <- "https://raw.githubusercontent.com/nytimes/covid-19-data/master/mask-use/mask-use-by-county.csv"
MaskUse <- read_csv(myfile)
```

```
## Rows: 3142 Columns: 6
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): COUNTYFP
## dbl (5): NEVER, RARELY, SOMETIMES, FREQUENTLY, ALWAYS
```

```
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# assigning new column name
colnames(MaskUse)[1] <- "fips"

fipsDF <- fips_info(MaskUse$fips)
head(fipsDF)
```

```
##     full abbr                      county  fips
## 1 Alaska   AK Aleutians West Census Area 02016
## 2 Alaska   AK     Aleutians East Borough 02013
## 3 Alaska   AK       Kusilvak Census Area 02158
## 4 Alaska   AK Lake and Peninsula Borough 02164
## 5 Alaska   AK  Matanuska-Susitna Borough 02170
## 6 Alaska   AK           Nome Census Area 02180
```

```
MaskUseDF <- merge(x = MaskUse, y = fipsDF, by = "fips", all = TRUE)
head(MaskUseDF)
```

```
##    fips NEVER RARELY SOMETIMES FREQUENTLY ALWAYS    full abbr         county
## 1 01001 0.053  0.074     0.134      0.295  0.444 Alabama   AL Autauga County
## 2 01003 0.083  0.059     0.098      0.323  0.436 Alabama   AL Baldwin County
## 3 01005 0.067  0.121     0.120      0.201  0.491 Alabama   AL Barbour County
## 4 01007 0.020  0.034     0.096      0.278  0.572 Alabama   AL    Bibb County
## 5 01009 0.053  0.114     0.180      0.194  0.459 Alabama   AL  Blount County
## 6 01011 0.031  0.040     0.144      0.286  0.500 Alabama   AL Bullock County
```

```
# assigning new column name
colnames(MaskUseDF)[7] <- "state"

MaskUseDF <- MaskUseDF[-8]

MaskUseDF <- MaskUseDF %>%
    mutate(county = str_remove_all(county, " County"))

MaskUseDF <- na.omit(MaskUseDF)

# write.csv(MaskUseDF,"MaskUseClean.csv", row.names = FALSE)
```

2. Exploratory Data Analysis

a. College Data

Read in file:

```
colleges <- read.csv("colleges_edited.csv")
```

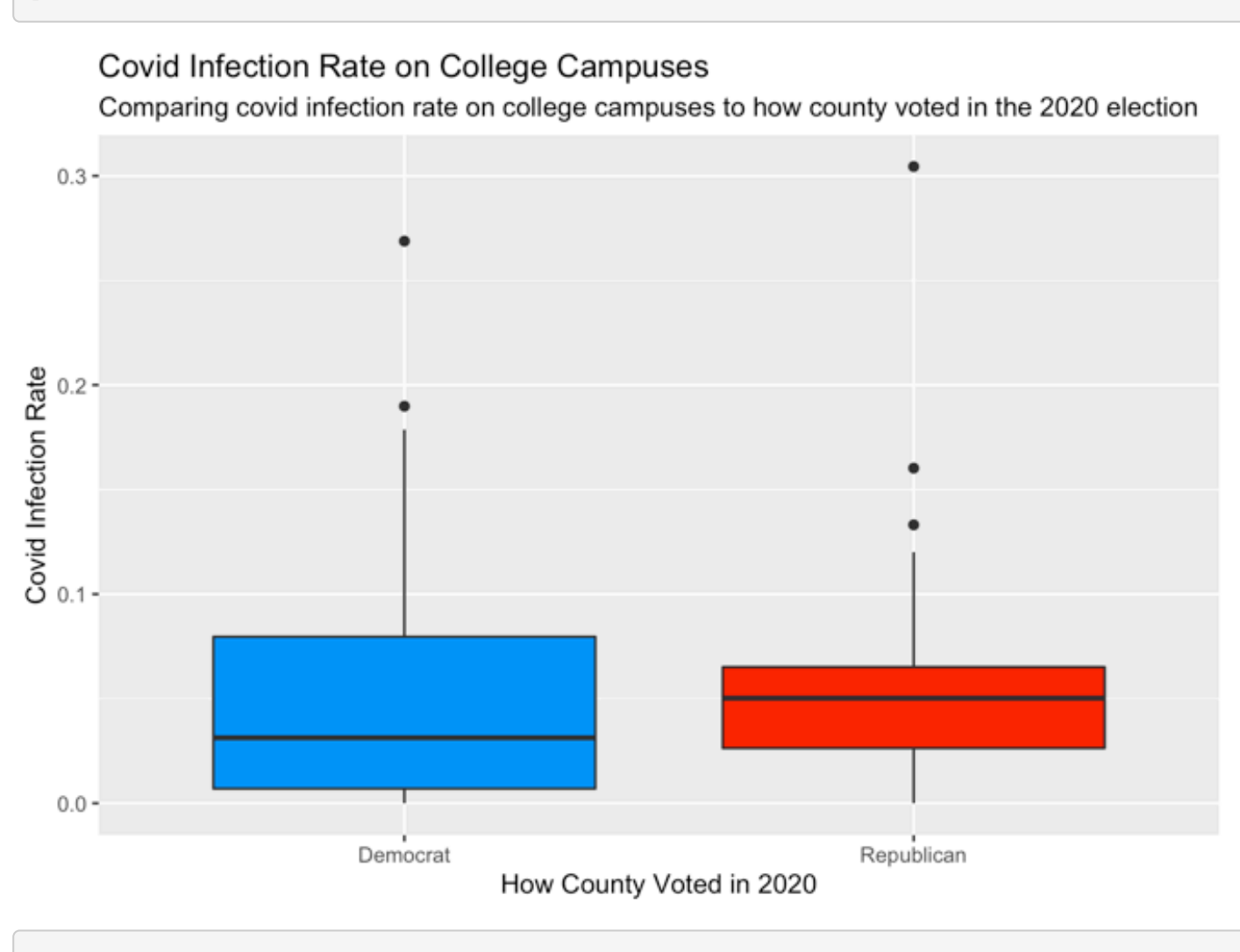
Conditional Probabilities for College data

```
print("Distribution of Covid case rates on college campuses: ")
summary(colleges$case_rate)

print("Distribution of the percent of county population that ALWAYS wears a mask,
      for counties with college campuses: ")
summary(colleges$ALWAYS)


## P(less than half of county with a college wears a mask | county voted Republican)
print(
  paste0(
    "P(less than half of county with a college wears a mask | county voted Republican) = ",
    nrow(colleges[(colleges$ALWAYS < 0.5) &
                  (colleges$winner == "Republican"),]) / nrow(colleges[(colleges$winner == "Republican"),])))


## P(low covid rate on campus (bottom 50 percentile) | county has high percent of mask wearers (top 50 percentile
))
print(
  paste0(
    "P(low covid rate on campus (bottom 50 percentile) | county has high percent of mask wearers (top 50 percenti
le)) = ",
    nrow(colleges[(colleges$ALWAYS > 0.6805) &
                  (colleges$case_rate < 0.04156),]) / nrow(colleges[(colleges$ALWAYS > 0.6805),])))


## P(low covid rate on campus (lower quantile) | county has high percent of mask wearers (upper quantile))
print(
  paste0(
    "P(low covid rate on campus (lower quantile) | county has high percent of mask wearers (upper quantile)) = ",
    nrow(colleges[(colleges$ALWAYS > 0.7560) &
                  (colleges$case_rate < 0.07609), ]) / nrow(colleges[(colleges$ALWAYS > 0.7560), ])))
```

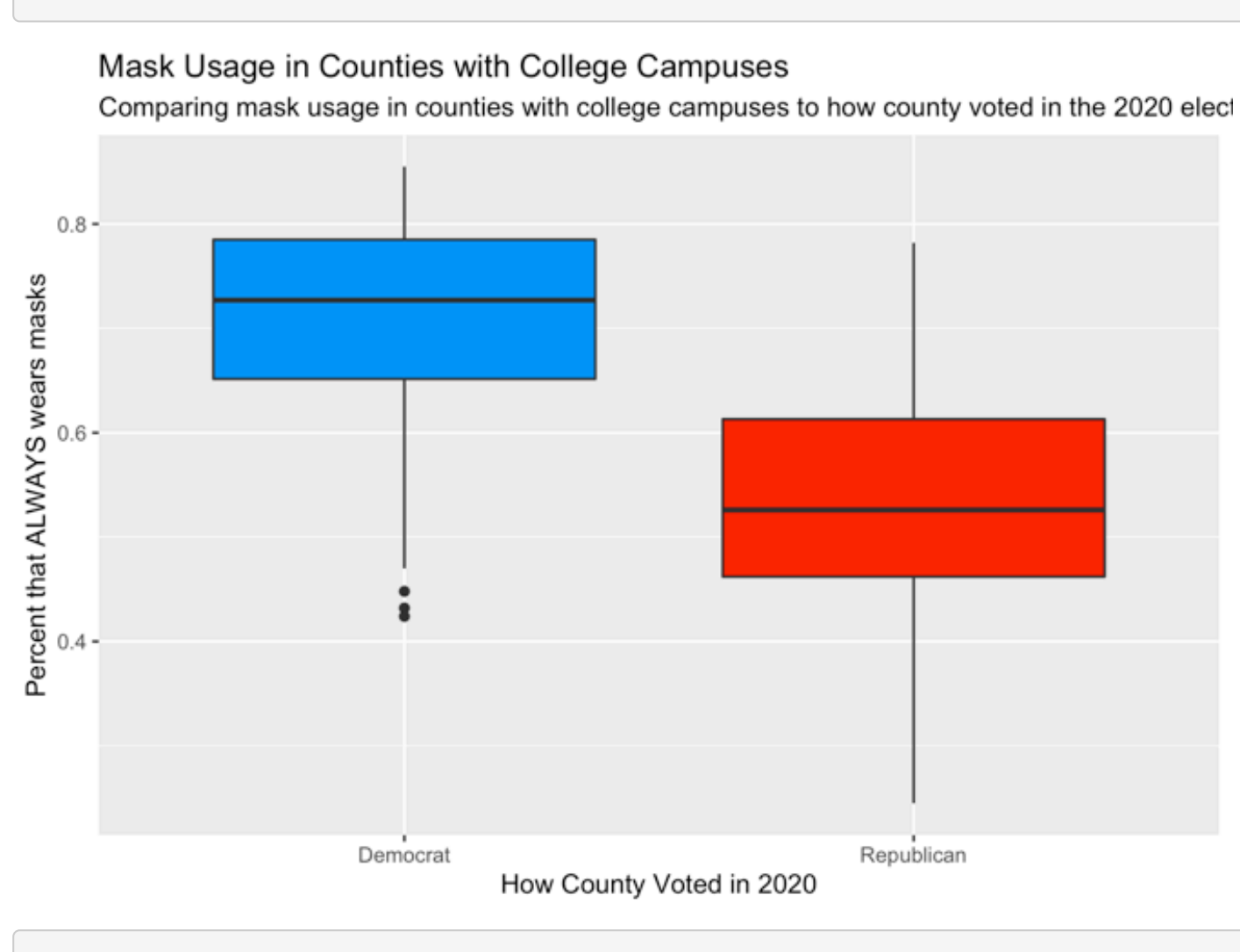
```
## [1] "Distribution of Covid case rates on college campuses: "
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00000 0.01302 0.04081 0.05057 0.07196 0.30448 
## [1] "Distribution of the percent of county population that ALWAYS wears a mask, \n      for counties with coll
ege campuses: "
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2450  0.5675  0.6805  0.6504  0.7560  0.8550 
## [1] "P(less than half of county with a college wears a mask | county voted Republican) = 0.394366197183099"
## [1] "P(low covid rate on campus (bottom 50 percentile) | county has high percent of mask wearers (top 50 perce
ntile)) = 0.660550458715596"
## [1] "P(low covid rate on campus (lower quantile) | county has high percent of mask wearers (upper quantile)) =
0.833333333333333"
```

Visuals for College Data

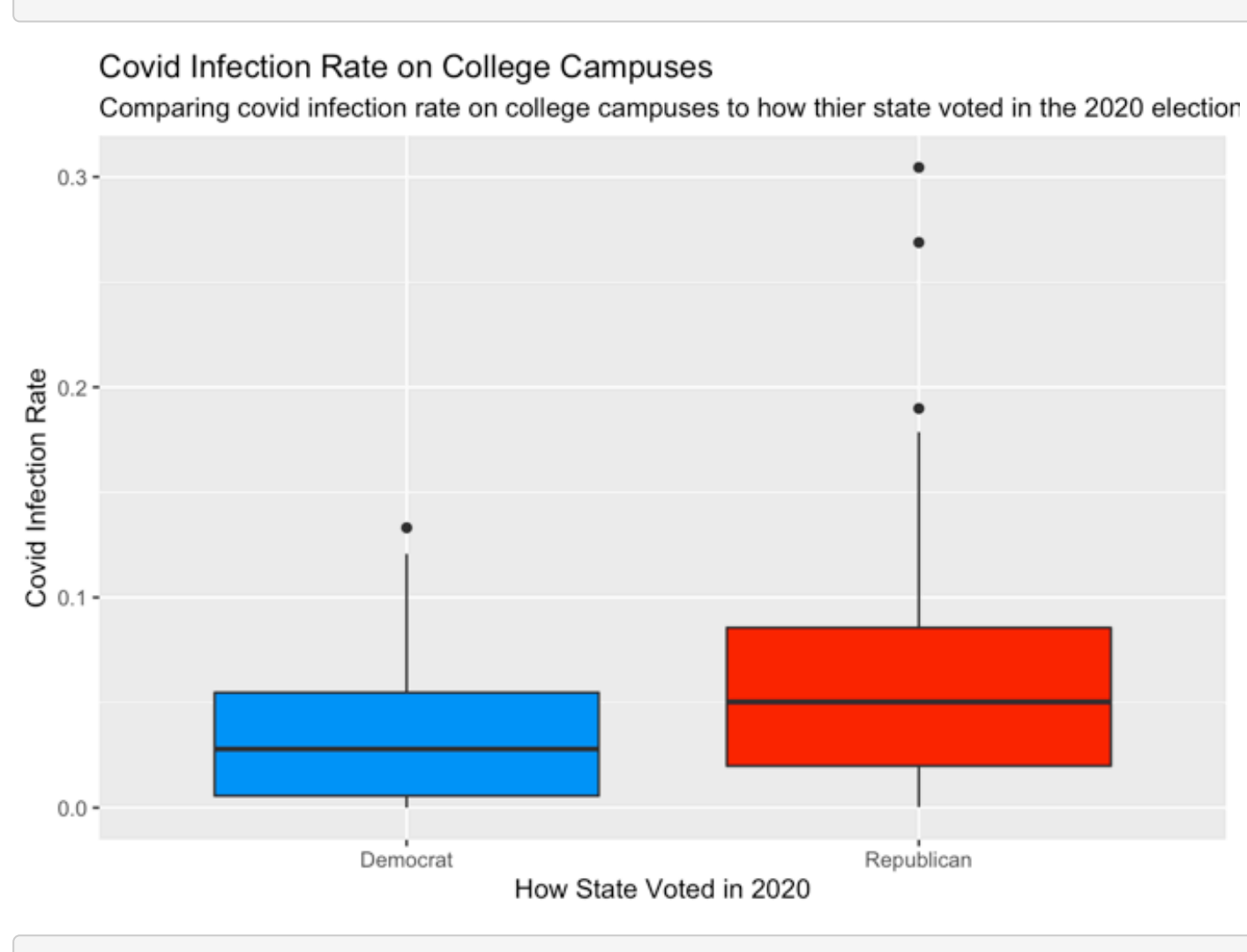
```
plot<-ggplot(colleges, aes(winner, case_rate)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "Covid Infection Rate on College Campuses",
       y = "Covid Infection Rate",
       x = "How County Voted in 2020",
       color = "Political Lean",
       subtitle = "Comparing covid infection rate on college campuses to how county voted in the 2020 election"
)

plot
```



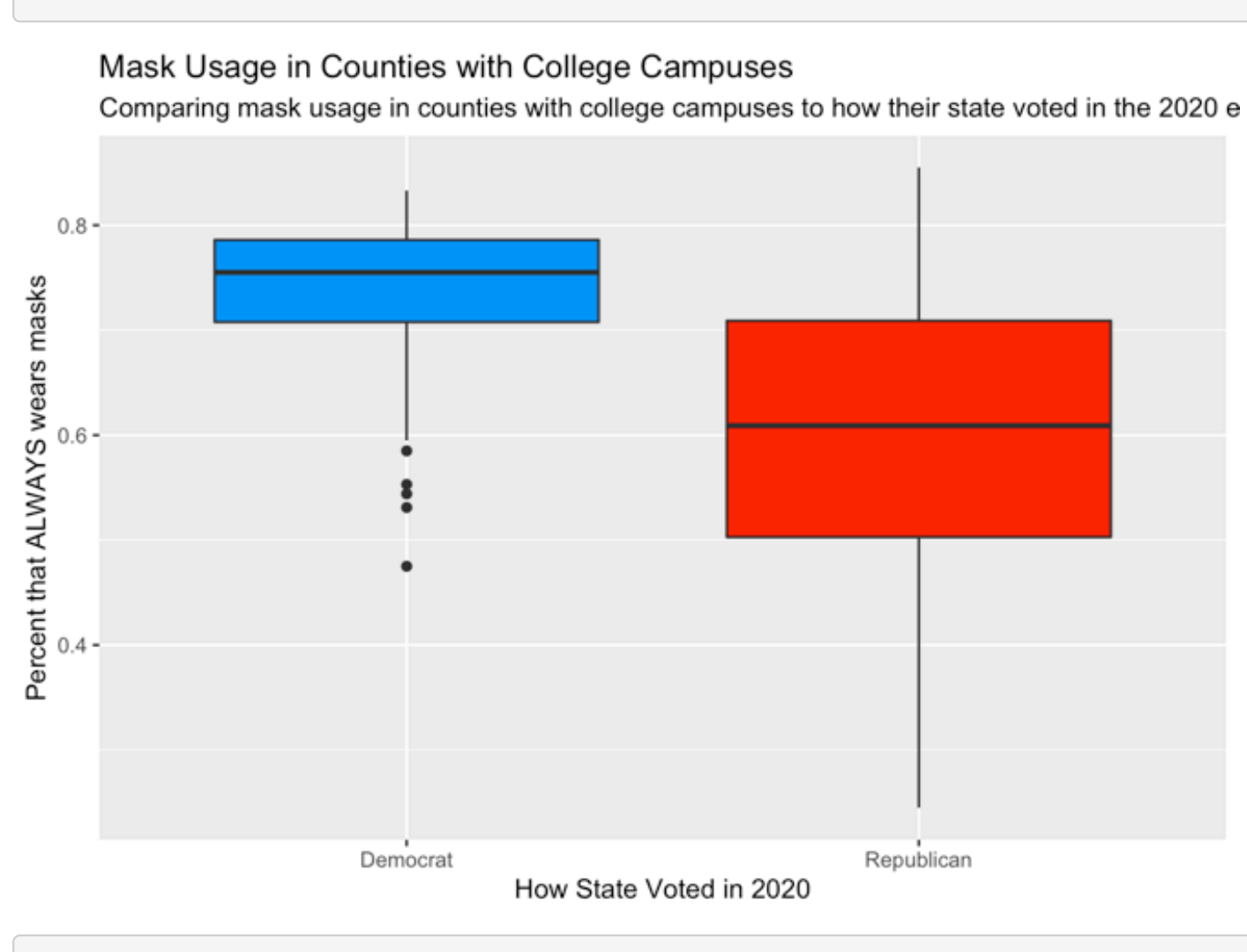
```
plot<-ggplot(colleges, aes(winner, ALWAYS)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "Mask Usage in Counties with College Campuses",
       y = "Percent that ALWAYS wears masks",
       x = "How County Voted in 2020",
       color = "Political Lean",
       subtitle = "Comparing mask usage in counties with college campuses to how county voted in the 2020 elect
ion")

plot
```



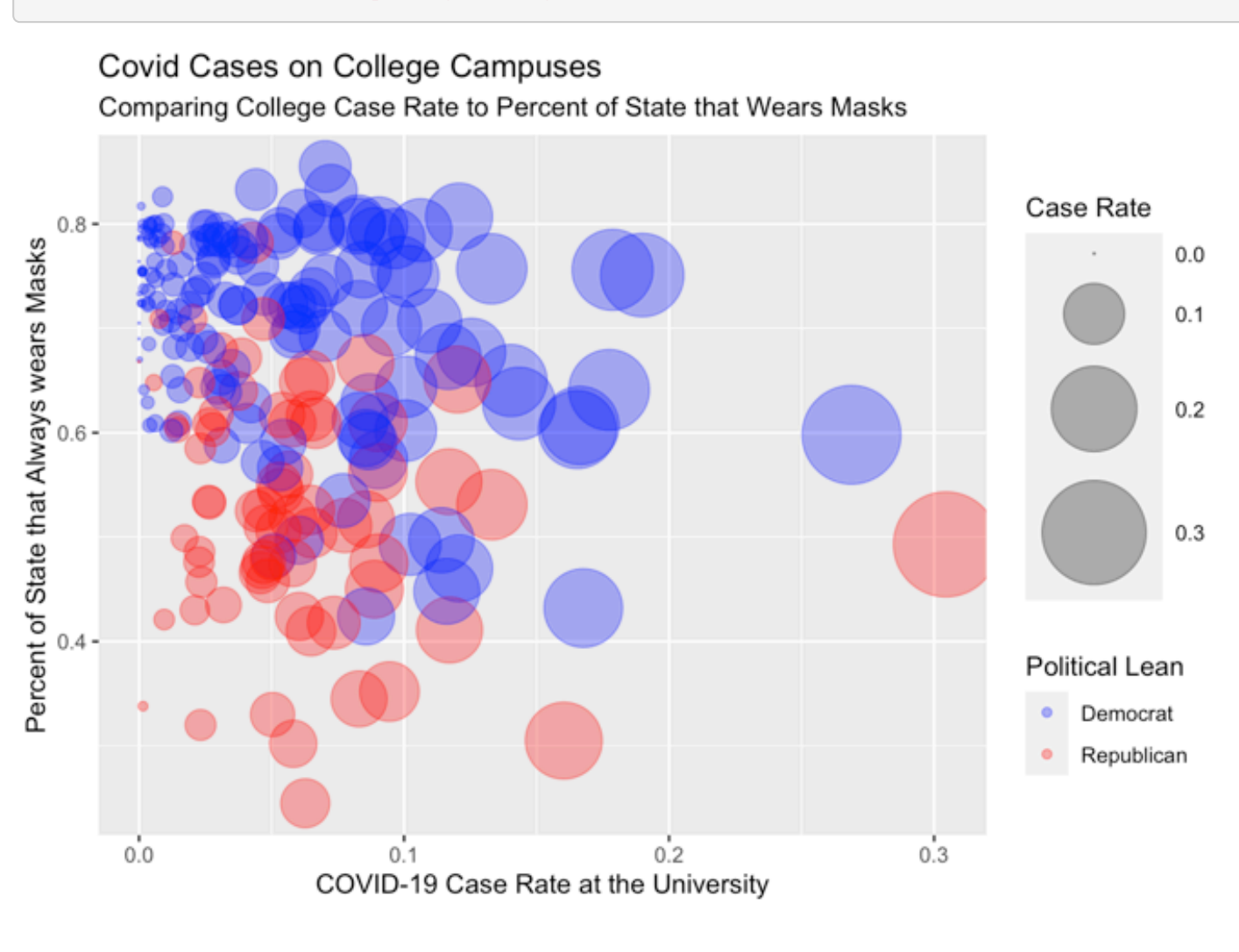
```
plot<-ggplot(colleges, aes(political_party, case_rate)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "Covid Infection Rate on College Campuses",
       y = "Covid Infection Rate",
       x = "How State Voted in 2020",
       color = "Political Lean",
       subtitle = "Comparing covid infection rate on college campuses to how thier state voted in the 2020 elec
tion")

plot
```



```
plot<-ggplot(colleges, aes(political_party, ALWAYS)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "Mask Usage in Counties with College Campuses",
       y = "Percent that ALWAYS wears masks",
       x = "How State Voted in 2020",
       color = "Political Lean",
       subtitle = "Comparing mask usage in counties with college campuses to how their state voted in the 2020
election")

plot
```



```
## scatter plot with different size points

colleges %>%
  arrange(desc(case_rate)) %>%
  mutate(winner = factor(winner)) %>%
  ggplot(aes(x=case_rate, y=ALWAYS, size=case_rate, color = winner)) +
    geom_point(alpha=0.35) +
    scale_size(range = c(.1, 20), name="Case Rate") +
    scale_color_manual(values = c("blue", "red")) +
    labs(title = "Covid Cases on College Campuses",
         y = "Percent of State that Always wears Masks",
         x = "COVID-19 Case Rate at the University",
         color = "Political Lean",
         subtitle = "Comparing College Case Rate to Percent of State that Wears Masks")
```



b. County Data

Read in file:

```
county_data <- read.csv("counties_edited.csv")
```

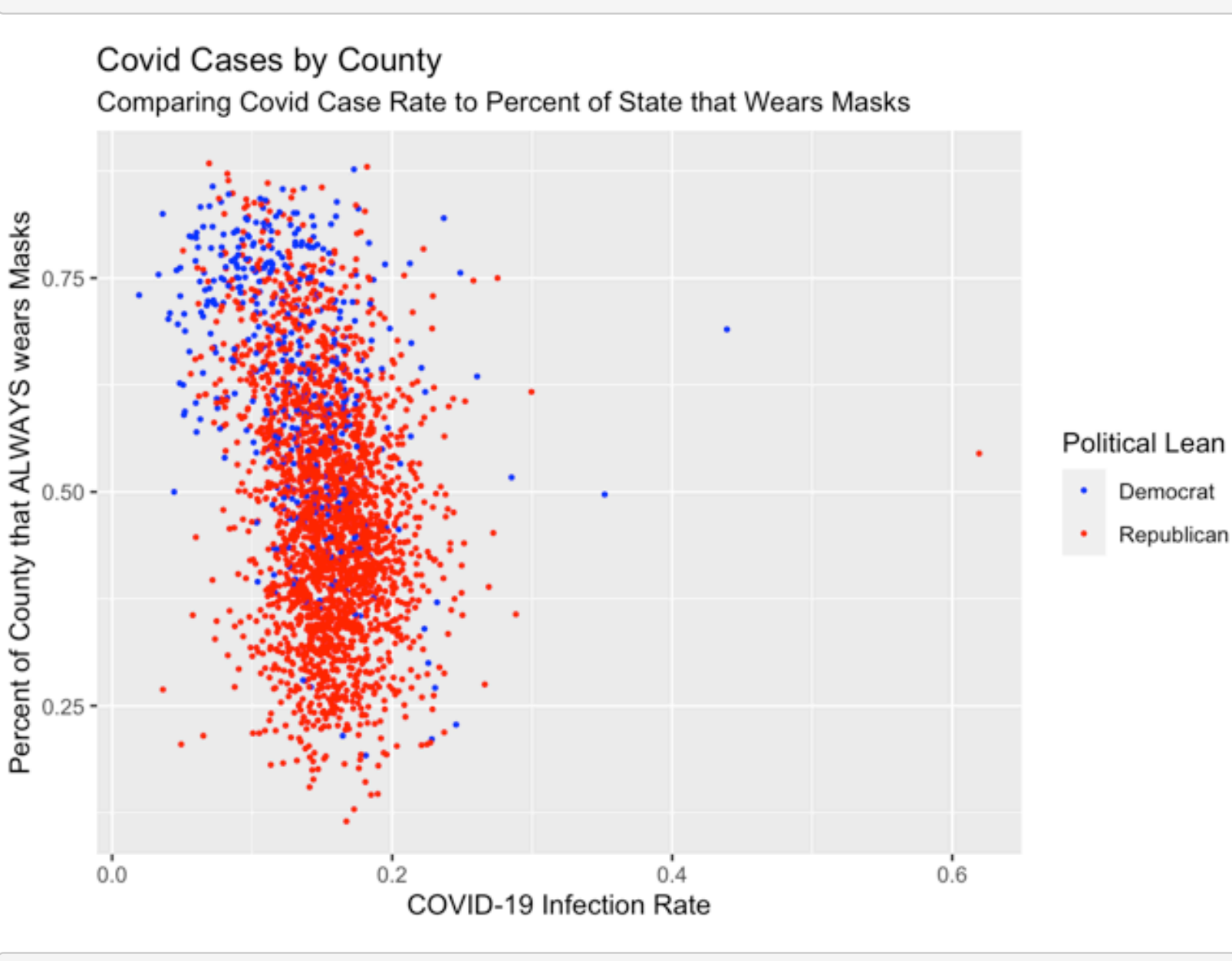
Conditional Probabilities for County data

```
print(paste0("Distribution of county populations that ALWAYS wear masks: "))
summary(county_data$ALWAYS)

## P(upper quartile of mask use | voted democrat)
print(paste0("P(upper quartile of mask use | voted democrat) = ",
nrow(county_data[(county_data$ALWAYS > 0.6080)&(county_data$winner == "Democrat"), ])/nrow(county_data[(county_da
ta$winner == "Democrat"), ])))


## P(upper HALF of mask use | voted Democrat)
print(paste0("P(upper HALF of mask use | voted Democrat) = ",
nrow(county_data[(county_data$ALWAYS > 0.5014)&(county_data$winner == "Democrat"), ])/nrow(county_data[(county_da
ta$winner == "Democrat"), ])))


## P(lower HALF of mask use | voted Republican)
print(paste0("P(lower HALF of mask use | voted Republican) = ",
             nrow(county_data[(county_data$ALWAYS < 0.5014)&(county_data$winner == "Republican"), ])
             /nrow(county_data[(county_data$winner == "Republican"), ])))


print(paste0("Distribution of county populations that NEVER wear masks: "))
summary(county_data$NEVER)

## P(upper quartile of NEVER mask use | voted democrat)
print(paste0("P(upper quartile of NEVER mask use | voted democrat) = ",
             nrow(county_data[(county_data$NEVER > 0.1180)&(county_data$winner == "Democrat"), ])
             /nrow(county_data[(county_data$winner == "Democrat"), ])))


## P(upper HALF of NEVER mask use | voted Democrat)
print(paste0("P(upper HALF of NEVER mask use | voted Democrat) = ",
             nrow(county_data[(county_data$NEVER > 0.0710)&(county_data$winner == "Democrat"), ])
             /nrow(county_data[(county_data$winner == "Democrat"), ])))


## P(lower HALF of NEVER mask use | voted Republican)
print(paste0("P(lower HALF of NEVER mask use | voted Republican) = ",
             nrow(county_data[(county_data$NEVER < 0.0710)&(county_data$winner == "Republican"), ])
             /nrow(county_data[(county_data$winner == "Republican"), ])))



## P(upper quartile of NEVER mask use | voted republican)
print(paste0("P(upper quartile of NEVER mask use | voted republican) = ",
             nrow(county_data[(county_data$NEVER > 0.1180)&(county_data$winner == "Republican"), ])
             /nrow(county_data[(county_data$winner == "Republican"), ])))


## P(upper HALF of NEVER mask use | voted republican)
print(paste0("P(upper HALF of NEVER mask use | county voted republican) = ",
             nrow(county_data[(county_data$NEVER > 0.0710)&(county_data$winner == "Republican"), ])
             /nrow(county_data[(county_data$winner == "Republican"), ])))


## P(lower HALF of NEVER mask use | voted Democrat)
print(paste0("P(lower HALF of NEVER mask use | county voted Democrat) = ", nrow(county_data[(county_data$NEVER <
0.0710)&(county_data$winner == "Democrat"), ])/nrow(county_data[(county_data$winner == "Democrat"), ])))
```

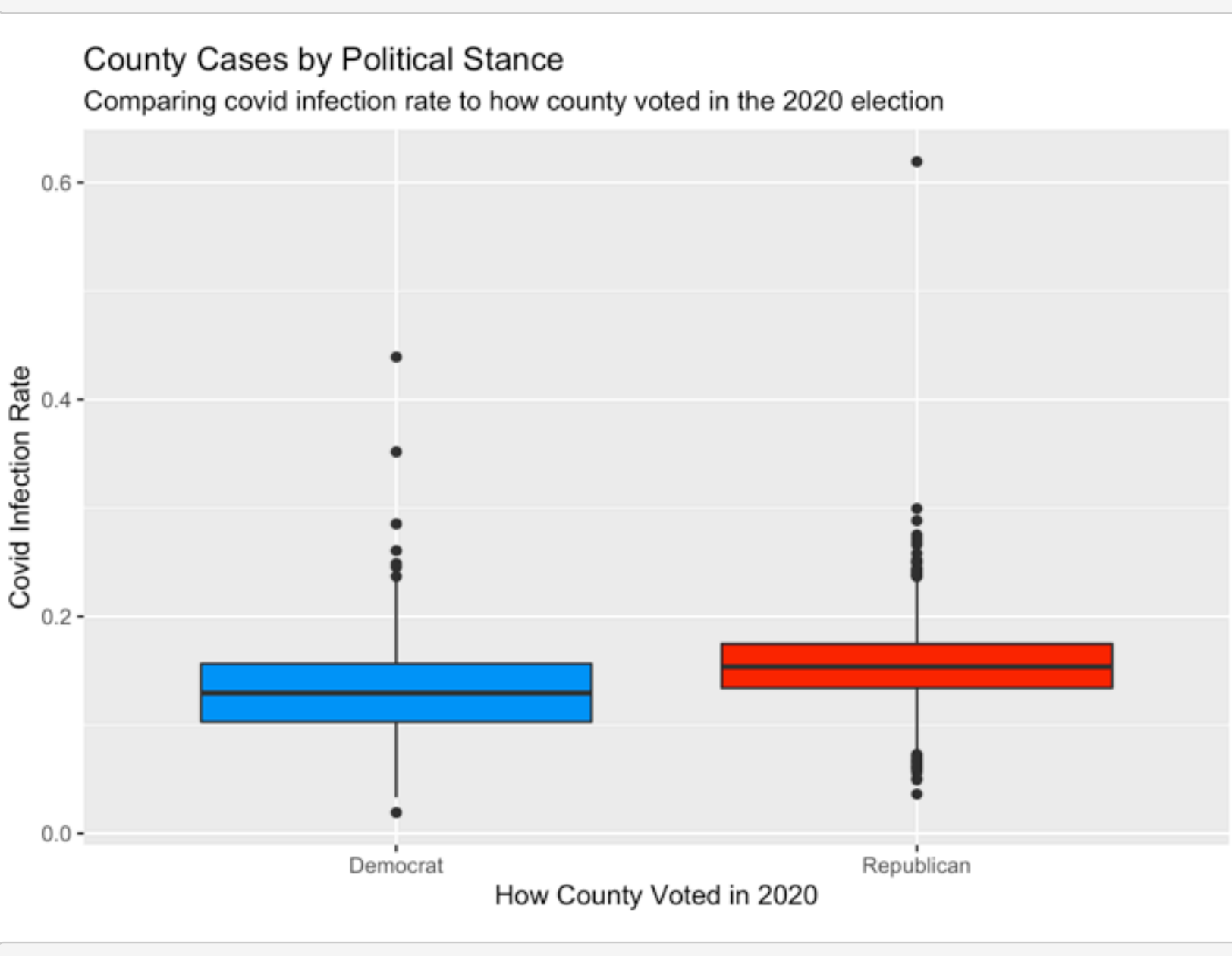
```
## [1] "Distribution of county populations that ALWAYS wear masks: "
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1150  0.3890  0.4920  0.5014  0.6080  0.8840
## [1] "P(upper quartile of mask use | voted democrat) = 0.623471882640587"
## [1] "P(upper HALF of mask use | voted Democrat) = 0.821515892420538"
## [1] "P(lower HALF of mask use | voted Republican) = 0.589100346020761"
## [1] "Distribution of county populations that NEVER wear masks: "
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0350  0.0710  0.0828  0.1180  0.4320
## [1] "P(upper quartile of NEVER mask use | voted democrat) = 0.0855745721271394"
## [1] "P(upper HALF of NEVER mask use | voted Democrat) = 0.198044009779951"
## [1] "P(lower HALF of NEVER mask use | voted Republican) = 0.444204152249135"
## [1] "P(upper quartile of NEVER mask use | voted republican) = 0.277249134948097"
## [1] "P(upper HALF of NEVER mask use | county voted republican) = 0.551470588235294"
## [1] "P(lower HALF of NEVER mask use | county voted Democrat) = 0.797066014669927"
```

Visuals for County data

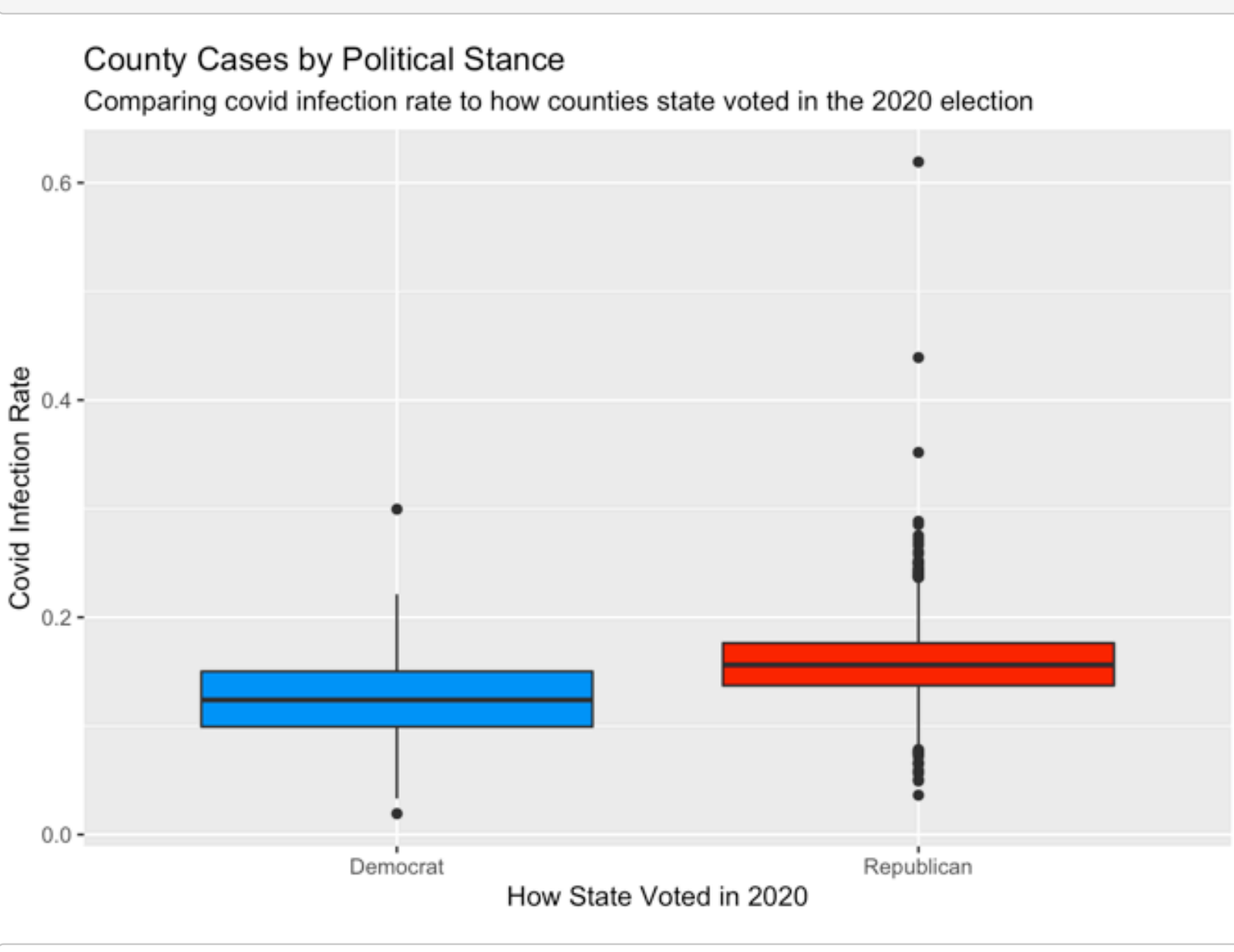
```
county_data %>%
  arrange(desc(case_rate)) %>%
  mutate(winner = factor(winner)) %>%
  ggplot(aes(x=case_rate, y=ALWAYS)) + #, size=case_rate, color = winner)) +
    geom_point(aes(color = winner), size = 0.5) +
    # scale_size(range = c(.05, 15), name="Case Rate") +
    scale_color_manual(values = c("blue", "red")) +
    labs(title = "Covid Cases by County",
         y = "Percent of County that ALWAYS wears Masks",
         x = "COVID-19 Infection Rate",
         color = "Political Lean",
         subtitle = "Comparing Covid Case Rate to Percent of State that Wears Masks")
```



```
summary(county_data[county_data$winner == "Republican", "case_rate"])
summary(county_data[county_data$winner == "Democrat", "case_rate"])


library(ggplot2)
plot<-ggplot(county_data, aes(winner, case_rate)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "County Cases by Political Stance",
         y = "Covid Infection Rate",
         x = "How County Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing covid infection rate to how county voted in the 2020 election")

plot
```

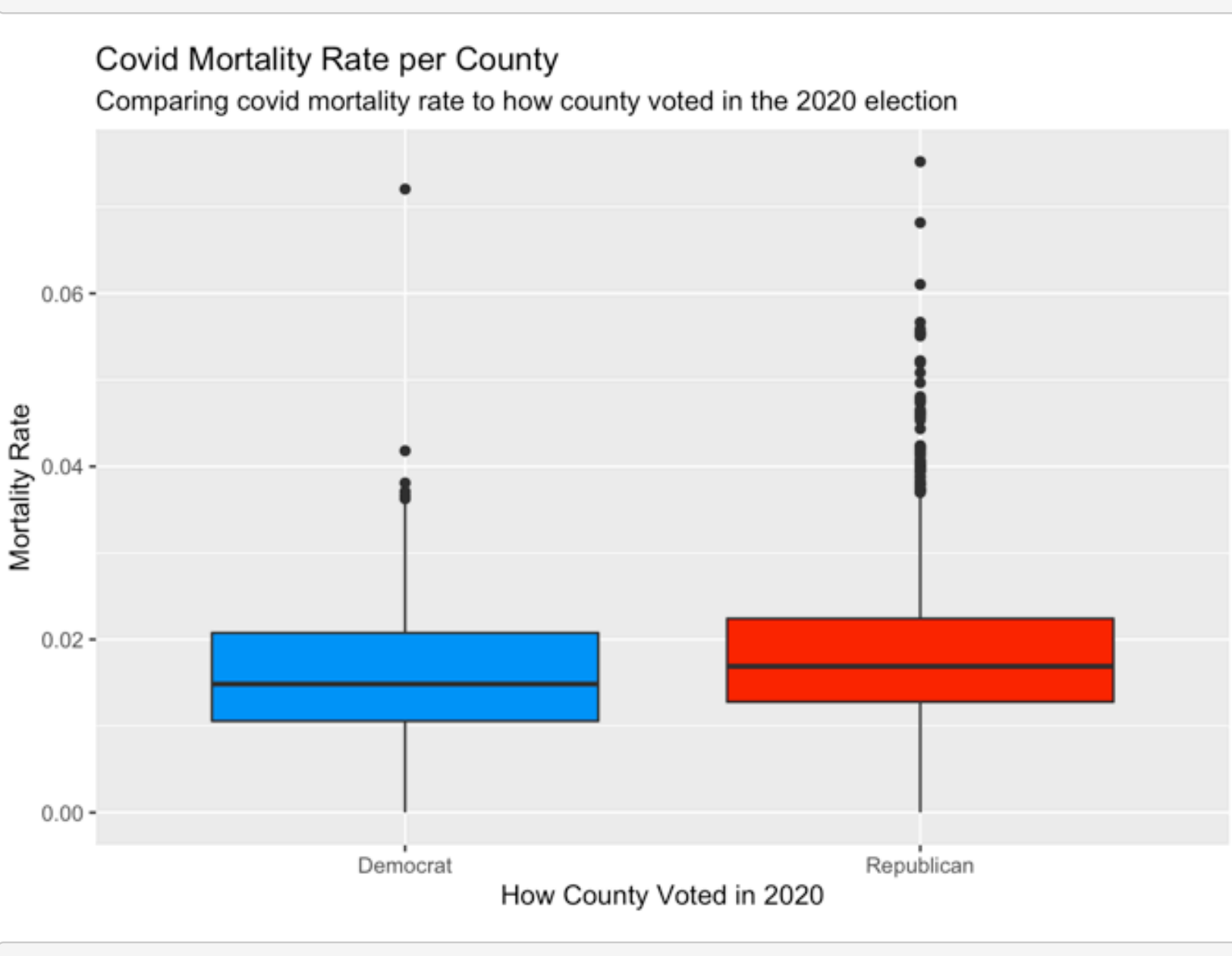


```
plot2<-ggplot(county_data, aes(political_party, case_rate)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "County Cases by Political Stance",
         y = "Covid Infection Rate",
         x = "How State Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing covid infection rate to how counties state voted in the 2020 election")

plot2
```

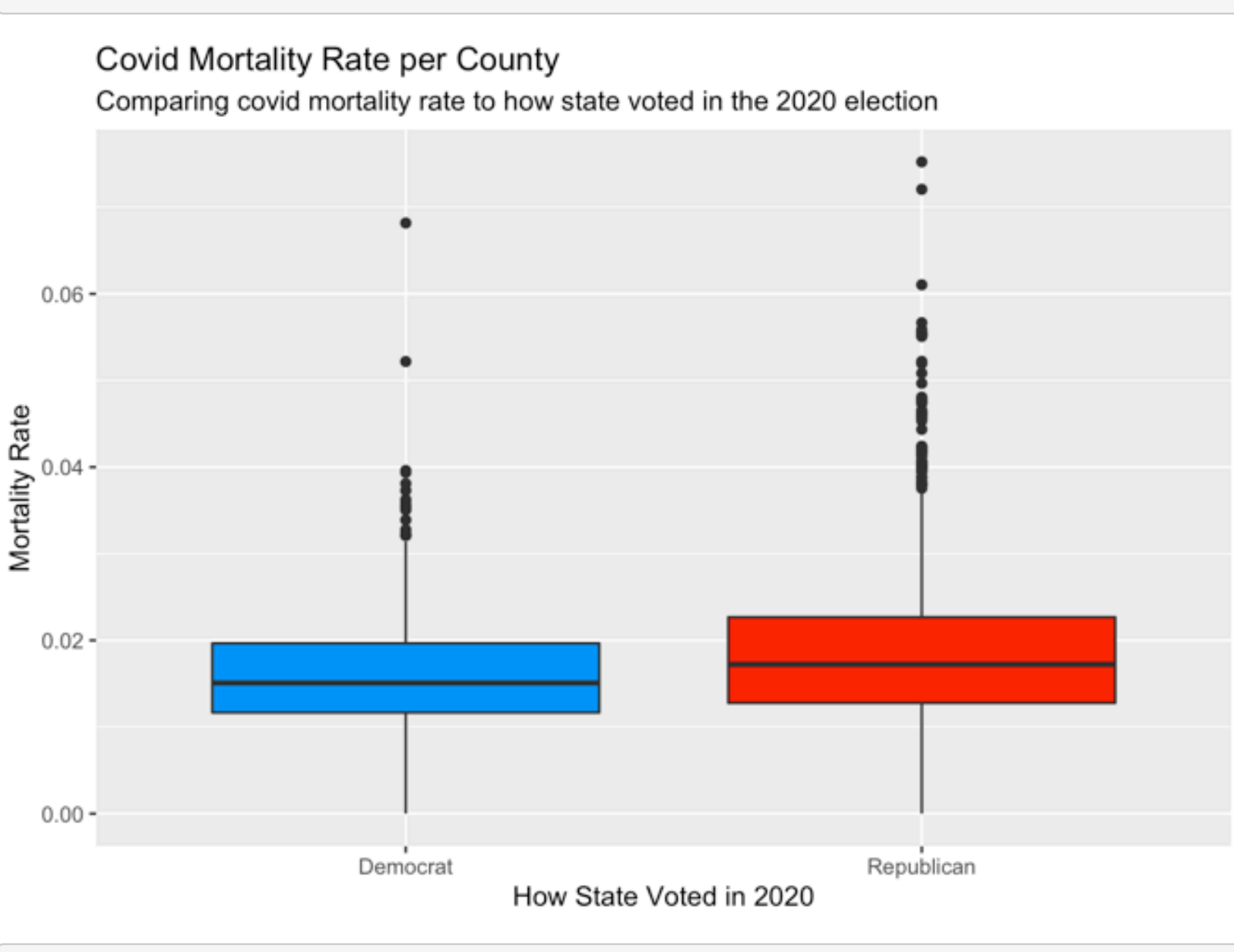


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03629 0.13410 0.15358 0.15402 0.17472 0.61924
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01928 0.10286 0.12939 0.12952 0.15643 0.43912
```

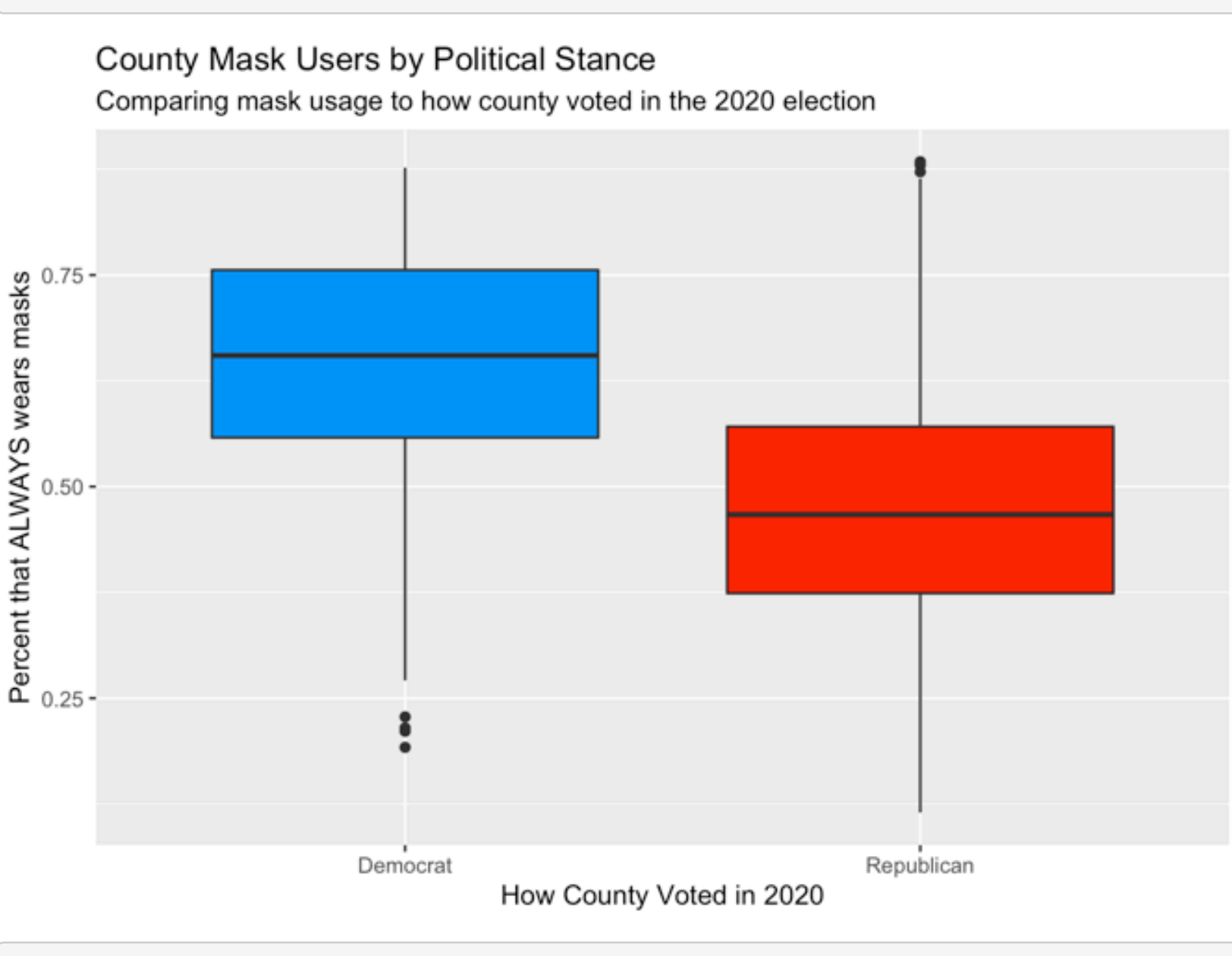
```
plot<-ggplot(county_data, aes(winner, percent_dead_from_covid)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "Covid Mortality Rate per County",
         y = "Mortality Rate",
         x = "How County Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing covid mortality rate to how county voted in the 2020 election")

plot
```



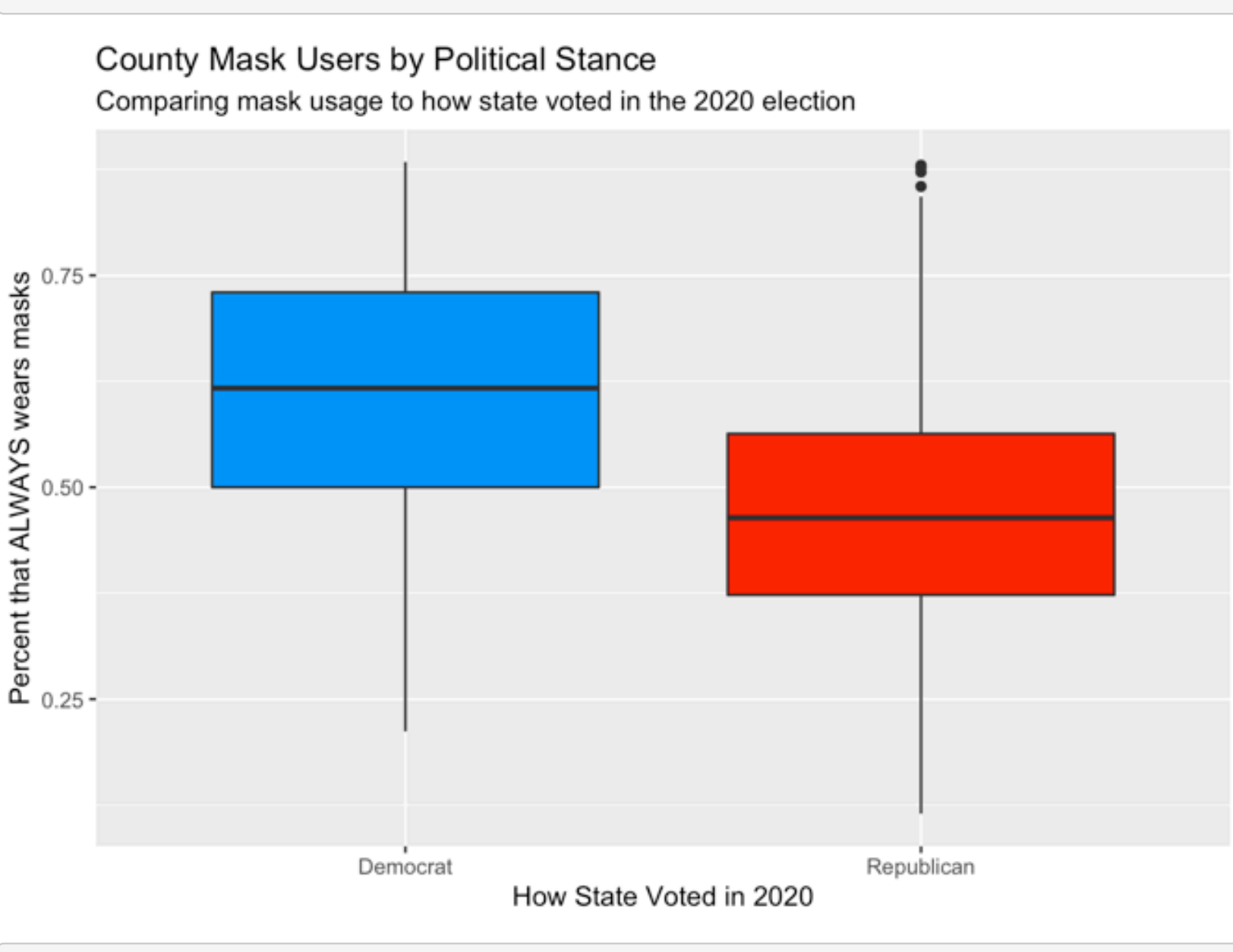
```
plot2<-ggplot(county_data, aes(political_party, percent_dead_from_covid)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "Covid Mortality Rate per County",
         y = "Mortality Rate",
         x = "How State Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing covid mortality rate to how state voted in the 2020 election")

plot2
```



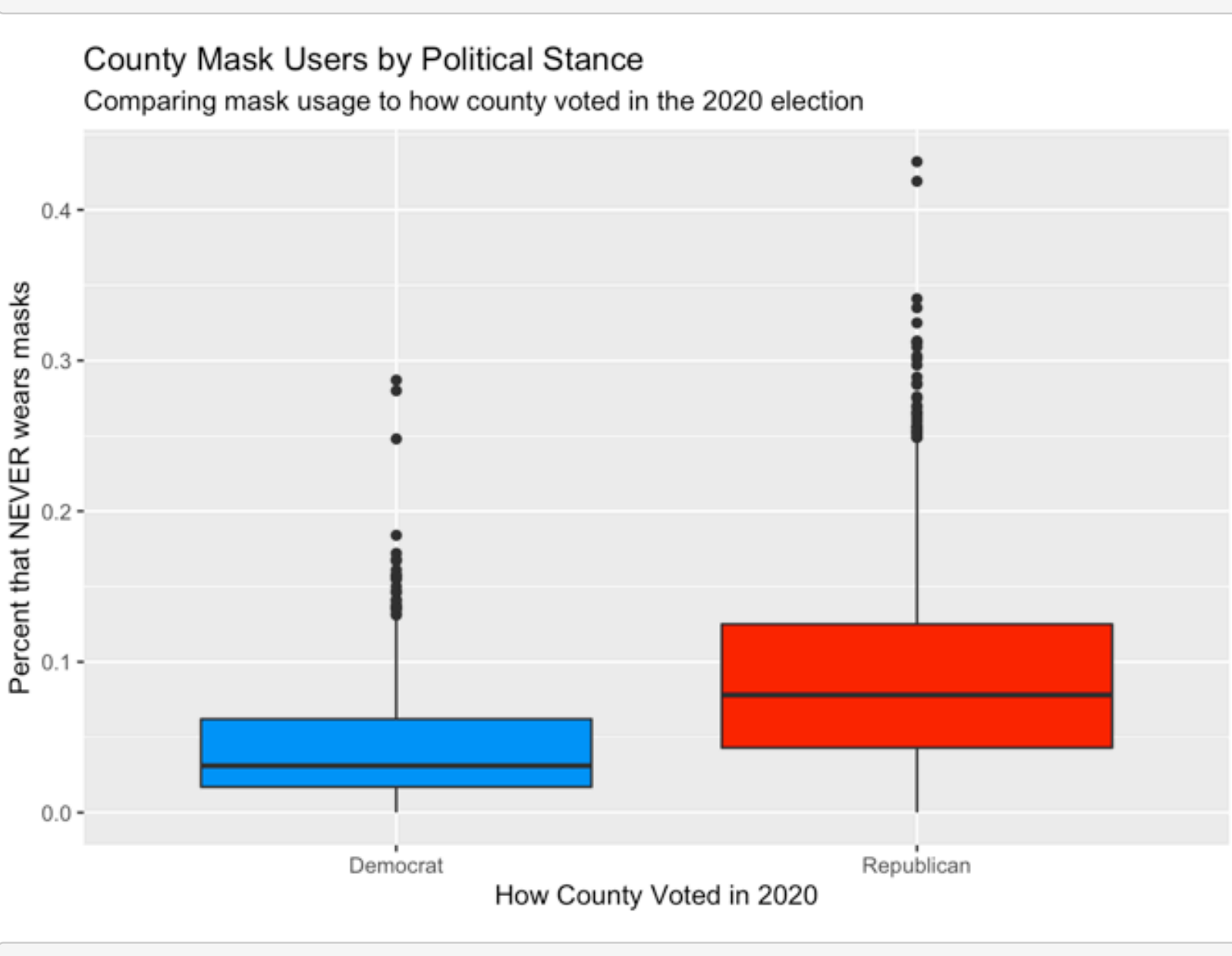
```
plot<-ggplot(county_data, aes(winner, ALWAYS)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "County Mask Users by Political Stance",
         y = "Percent that ALWAYS wears masks",
         x = "How County Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing mask usage to how county voted in the 2020 election")

plot
```



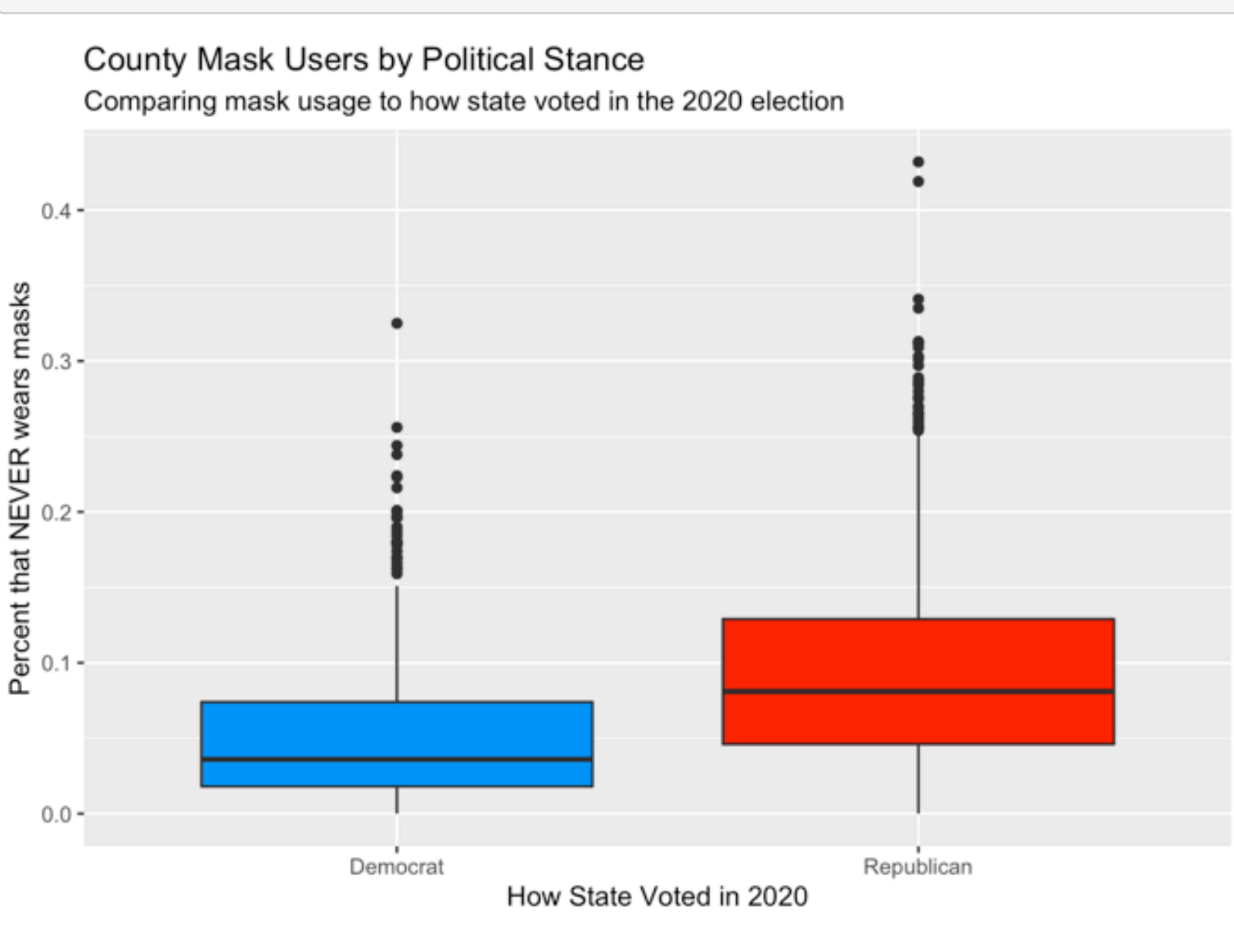
```
plot2<-ggplot(county_data, aes(political_party, ALWAYS)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "County Mask Users by Political Stance",
         y = "Percent that ALWAYS wears masks",
         x = "How State Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing mask usage to how state voted in the 2020 election")
plot2
```



```
plot<-ggplot(county_data, aes(winner, NEVER)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  labs(title = "County Mask Users by Political Stance",
         y = "Percent that NEVER wears masks",
         x = "How County Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing mask usage to how county voted in the 2020 election")

plot
```



```
plot2<-ggplot(county_data, aes(political_party, NEVER)) +
  geom_boxplot(fill = c("#0096fa", "#fa1100")) +
  scale_x_discrete(labels = c('Democrat', 'Republican')) +
  labs(title = "County Mask Users by Political Stance",
         y = "Percent that NEVER wears masks",
         x = "How State Voted in 2020",
         color = "Political Lean",
         subtitle = "Comparing mask usage to how state voted in the 2020 election")
plot2
```



3. Hypothesis Testing

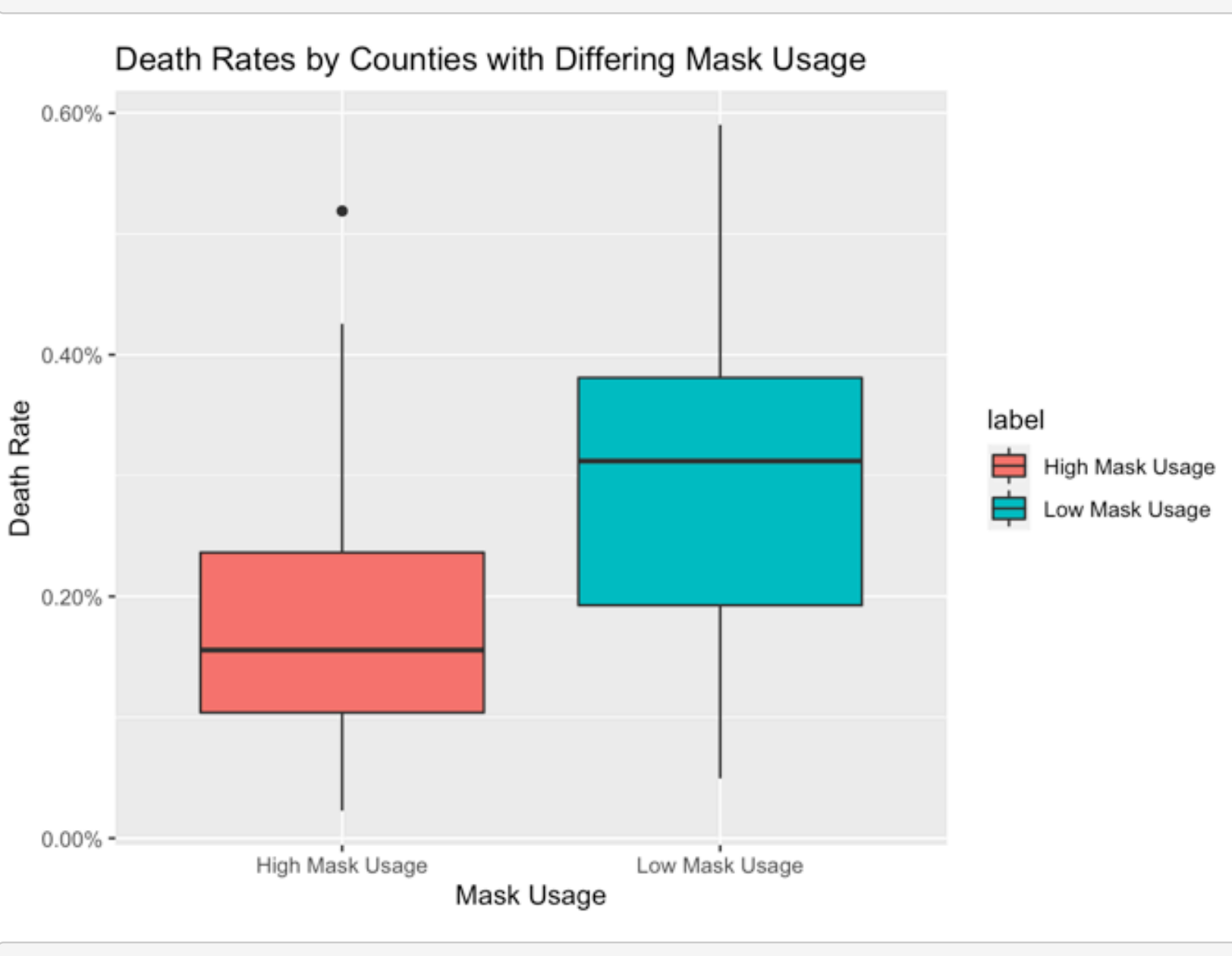
```
#Uploading data, cleaning county census data population estimates & mask data
#-------------------------------------------------------------------------------
countyDeaths <- read.csv('../live/us-counties.csv')#deaths by county dataset
countyPopulations <- read.csv('PopulationEstimates_csv.csv')#populations by county dataset
masks <- read.csv('../mask-use/mask-use-by-county.csv')#mask usage dataset
names <- countyPopulations[1, ]
names[1] <- "fips" #Changing name for merging
names[1,8] <- "Population"
countyPopulations <- countyPopulations[-c(1:2), ]
names(countyPopulations) <- names
countyPopulations$fips <- as.numeric(countyPopulations$fips) #converting fips codes to numeric for converting
```

```
## Warning: NAs introduced by coercion
```

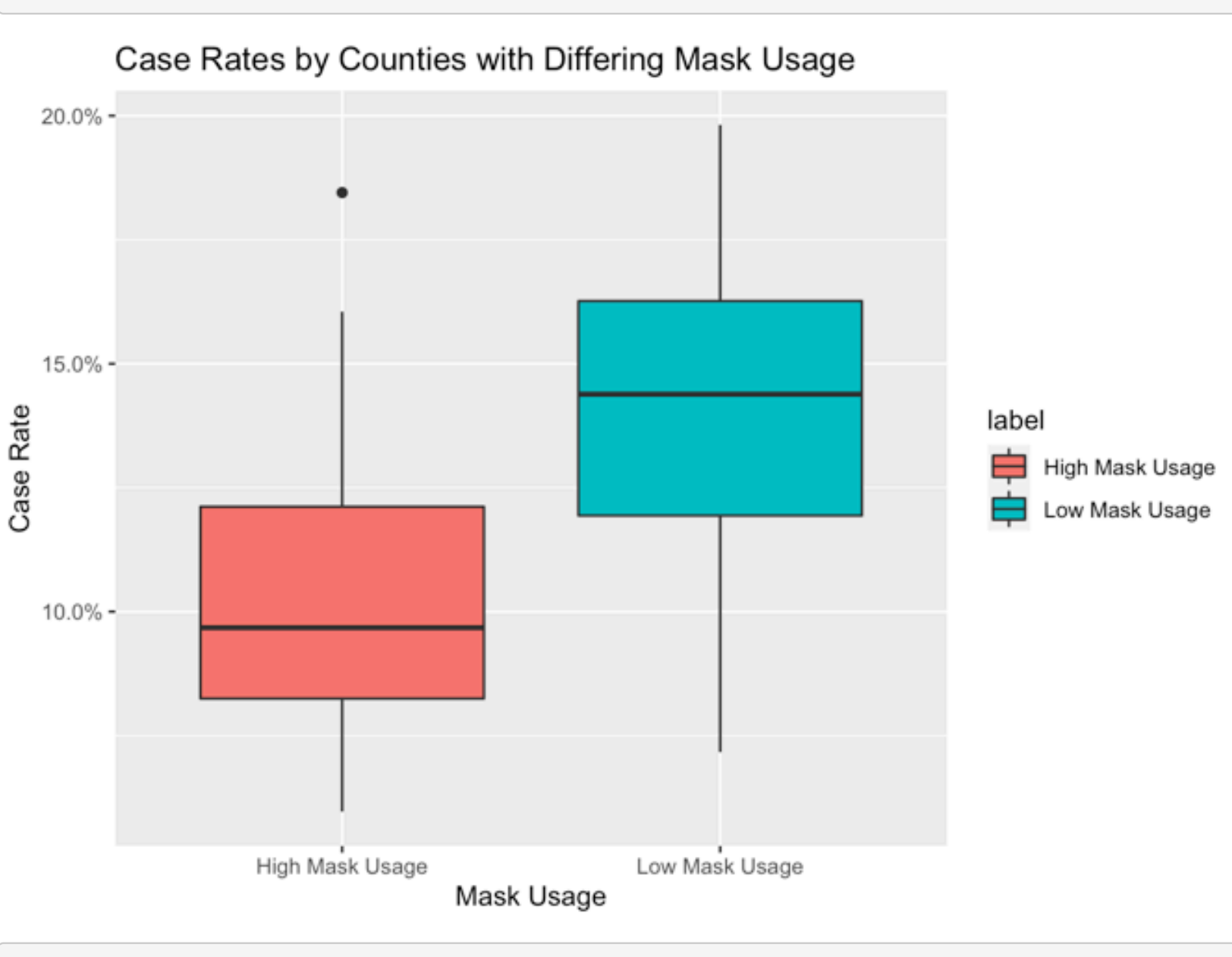
```
maskNames <- names(masks)
maskNames[1] <- "fips" #Changing name for merging
names(masks) <- maskNames

#Merging based on fips code, selecting key variables, cleaning resulting dataset
#-------------------------------------------------------------------------------
df <- merge(countyPopulations, countyDeaths, by = "fips" ) %>%
  select ('fips', 'state', 'Area name', 'Population', 'confirmed_cases', 'confirmed_deaths')
```

```
## Warning in merge.data.frame(countyPopulations, countyDeaths, by = "fips"):
## column names 'NA', 'NA', 'NA', 'NA', 'NA' are duplicated in the result
```

```
df$Population <- gsub("\\,", "", df$Population) #removing commas from population data
df$Population <- as.numeric(df$Population)
df$confirmed_deaths <- as.numeric(df$confirmed_deaths)
df$confirmed_cases <- as.numeric(df$confirmed_cases)
df <- df %>% mutate(case_rate = confirmed_cases/Population, death_rate = confirmed_deaths/Population) #calcualtin
g case rate
newdf <- merge(df, masks, by = "fips")
noMaskCounties <- newdf %>% arrange(desc(NEVER, RARELY)) %>% drop_na() %>% slice(1:30) %>% mutate(label = "Low Ma
sk Usage") #Selecting 30 lowest mask using counties with sufficient data
maskCounties <- newdf %>% arrange(desc(ALWAYS, FREQUENTLY)) %>% drop_na() %>% slice(1:30) %>% mutate(label = "Hig
h Mask Usage")#Selecting 30 highest mask using counties with sufficient data
finalDF <- rbind(noMaskCounties, maskCounties)
View(finalDF)

#Creating visualizations
#-------------------------------------------------------------------------------
finalDF %>% ggplot(aes(x = label, y = death_rate, fill = label)) + geom_boxplot() +
  labs(title = "Death Rates by Counties with Differing Mask Usage", x = "Mask Usage", y = "Death Rate") +
  scale_y_continuous(labels = scales::percent)
```



```
finalDF %>% ggplot(aes(x = label, y = case_rate, fill = label)) + geom_boxplot() +
  labs(title = "Case Rates by Counties with Differing Mask Usage", x = "Mask Usage", y = "Case Rate") +
  scale_y_continuous(labels = scales::percent)
```



```
#T-Test
#-------------------------------------------------------------------------------
freqMaskDR <- subset(finalDF, select=death_rate, subset=label=="High Mask Usage", drop=T)
rarelyMaskDR <- subset(finalDF, select=death_rate, subset=label=="Low Mask Usage", drop=T)
t.test(rarelyMaskDR, freqMaskDR, alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  rarelyMaskDR and freqMaskDR
## t = 3.5594, df = 56.022, p-value = 0.000383
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0006240341          Inf
## sample estimates:
##   mean of x   mean of y
## 0.003031938 0.001854760
```

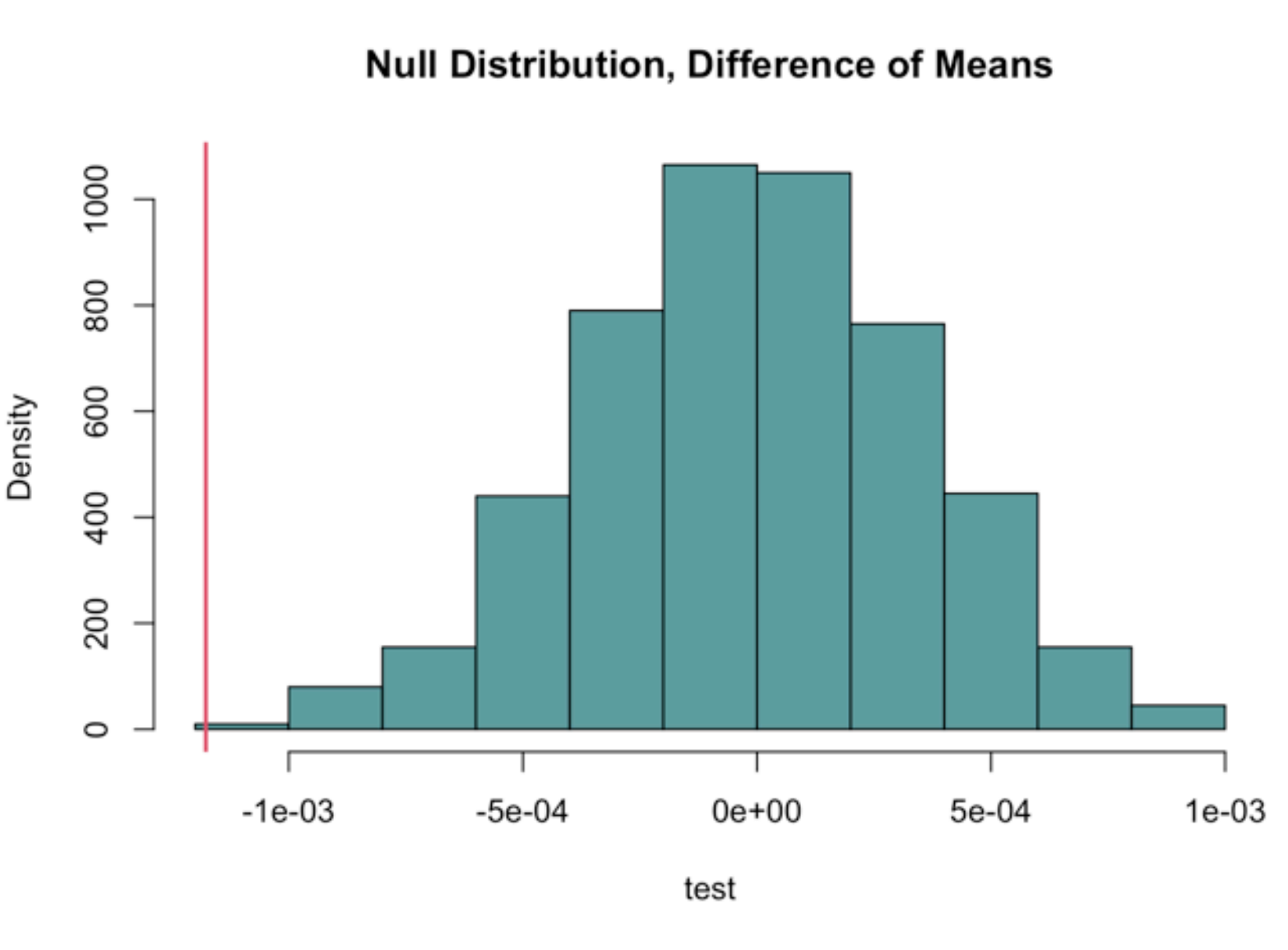
```
freqMaskCR <- subset(finalDF, select=case_rate, subset=label=="High Mask Usage", drop=T)
rarelyMaskCR <- subset(finalDF, select=case_rate, subset=label=="Low Mask Usage", drop=T)
t.test(rarelyMaskCR, freqMaskCR, alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  rarelyMaskCR and freqMaskCR
## t = 4.3449, df = 55.968, p-value = 2.963e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.02205518        Inf
## sample estimates:
## mean of x mean of y
## 0.1393027 0.1034439
```

```
#CHI-SQR Test
#-------------------------------------------------------------------------------
#DEATH RATE:
chisqDF <- finalDF %>% select(label, death_rate)

diffMean = function(chisqDF) {
  agg = aggregate(death_rate ~ label, data = chisqDF, FUN = mean)
  return(agg$death_rate[1] - agg$death_rate[2]) #xbar_c - xbar_t
}
myPerm <- function(){
  dfCopy = chisqDF
  dfCopy$death_rate <- dfCopy$death_rate[sample(60,60,replace=F)]
  diffMean(dfCopy)
}
stat = diffMean(chisqDF)
test = replicate(1000, myPerm())
mean(test < stat) #p-value
```

```
## [1] 0
```

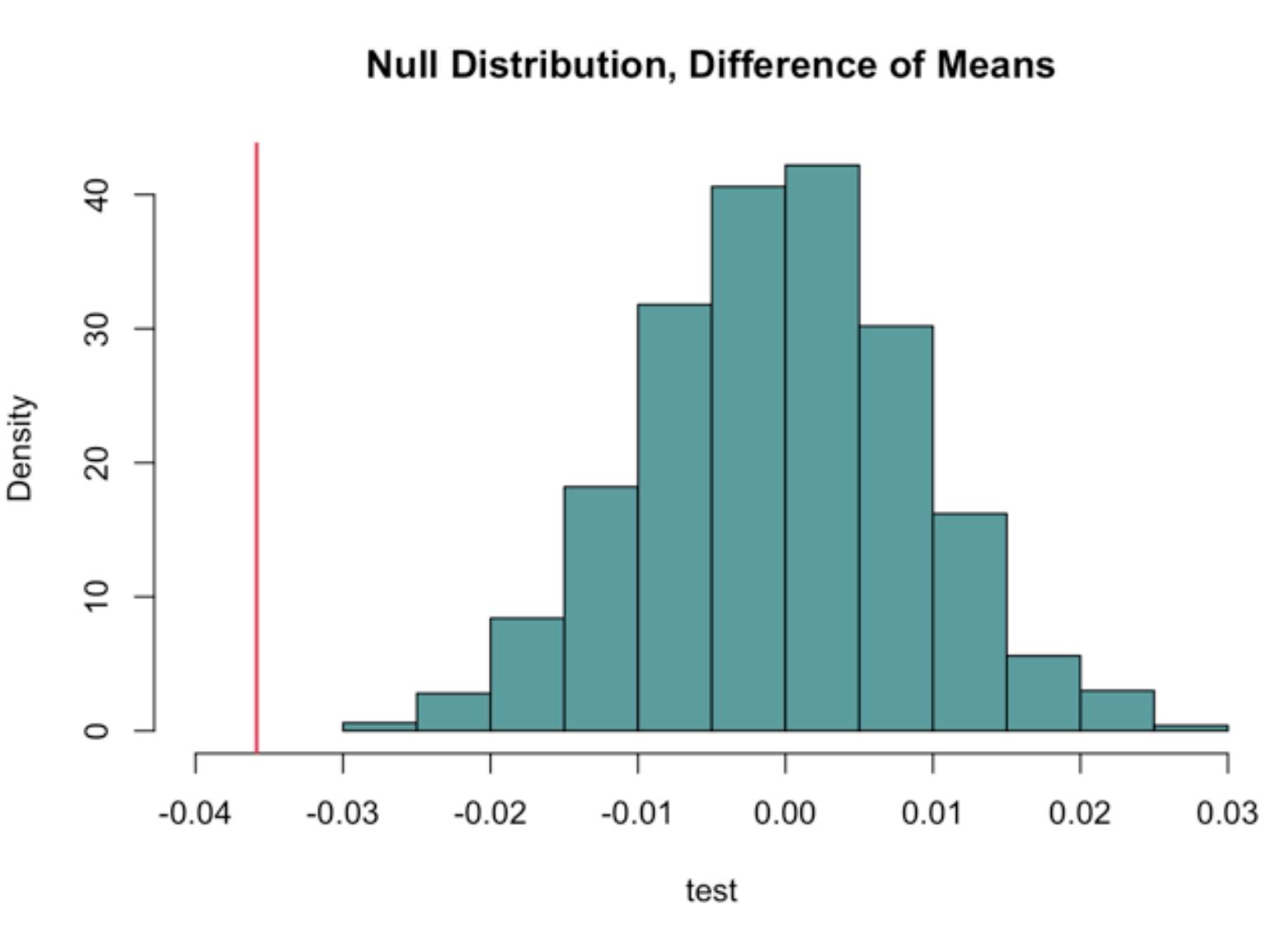
```
#Frequency distribution visualization
hist(test, main = "Null Distribution, Difference of Means", prob = T, col = "cadetblue")
abline(v = stat,col = 2, lwd = 2)
```



```
#----

#CASE RATE:
chisqDF <- finalDF %>% select(label, case_rate)

diffMean = function(chisqDF) {
  agg = aggregate(case_rate ~ label, data = chisqDF, FUN = mean)
  return(agg$case_rate[1] - agg$case_rate[2]) #xbar_c - xbar_t
}
myPerm <- function(){
  dfCopy = chisqDF
  dfCopy$case_rate <- dfCopy$case_rate[sample(60,60,replace=F)]
  diffMean(dfCopy)
}
stat = diffMean(chisqDF)
test = replicate(1000, myPerm())
mean(test < stat) #p-value
```

```
## [1] 0
```

```
#Frequency distribution visualization
range <- c(-0.04, 0.03)
hist(test, main = "Null Distribution, Difference of Means", prob = T, col = "cadetblue", xlim = range)
abline(v = stat,col = 2, lwd = 2)
```



Final Report

Christopher Delaney, Austin Eaton, Clara Richter, Elise Rust

12/8/2021

Load in necessary libraries

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
```

1. Colleges

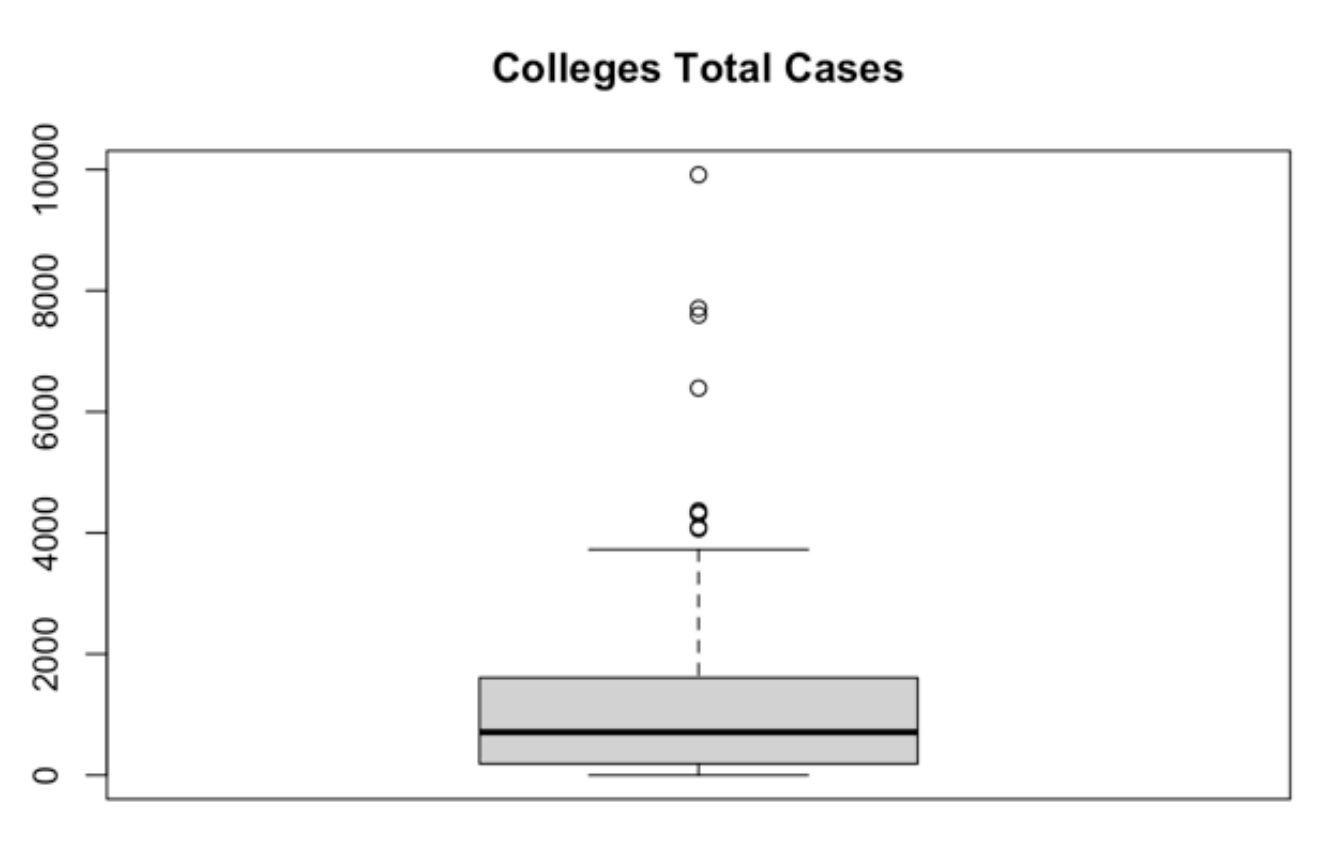
```
# Load in data
colleges_full <- read.csv("~/Desktop/Papers/Georgetown/Fall 2021/ANLY 511/Final Project/colleges_cleaned2.csv")

# Examine
head(colleges_full)
```

```
##   X   state                 county       city ipeds_id
## 1 1 Alabama                    Lee     Auburn   100858
## 2 2 Alabama                   Pike       Troy   102368
## 3 3 Alabama              Jefferson Birmingham   100663
## 4 4 Alabama                 Mobile     Mobile   102094
## 5 5  Alaska Anchorage Municipality  Anchorage   102553
## 6 6 Arizona               Maricopa    Phoenix   104717
##                              college cases rank2020 rank2015 enroll2020
## 1                  Auburn University  2742      101      150      30440
## 2                    Troy University   855      285      217      16981
## 3 University of Alabama at Birmingham  3143      193      256      21923
## 4        University of South Alabama   780      338      346      14834
## 5     University of Alaska Anchorage   186      348      283      14644
## 6            Grand Canyon University  1935        5       11      90253
##   enroll2015 NEVER RARELY SOMETIMES FREQUENTLY ALWAYS     winner
## 1      24864 0.053  0.064     0.138      0.183  0.562 Republican
## 2      20573 0.060  0.131     0.145      0.333  0.330 Republican
## 3      18568 0.049  0.037     0.107      0.179  0.628   Democrat
## 4      15065 0.077  0.070     0.102      0.244  0.506 Republican
## 5      17363 0.042  0.050     0.049      0.196  0.663       <NA>
## 6      55497 0.023  0.025     0.059      0.158  0.734   Democrat
##   perc_enroll_change  case_rate
## 1         0.22425997 0.09007884
## 2        -0.17459777 0.05035039
## 3         0.18068720 0.14336542
## 4        -0.01533355 0.05258191
## 5        -0.15659736 0.01270145
## 6         0.62626809 0.02143973
```

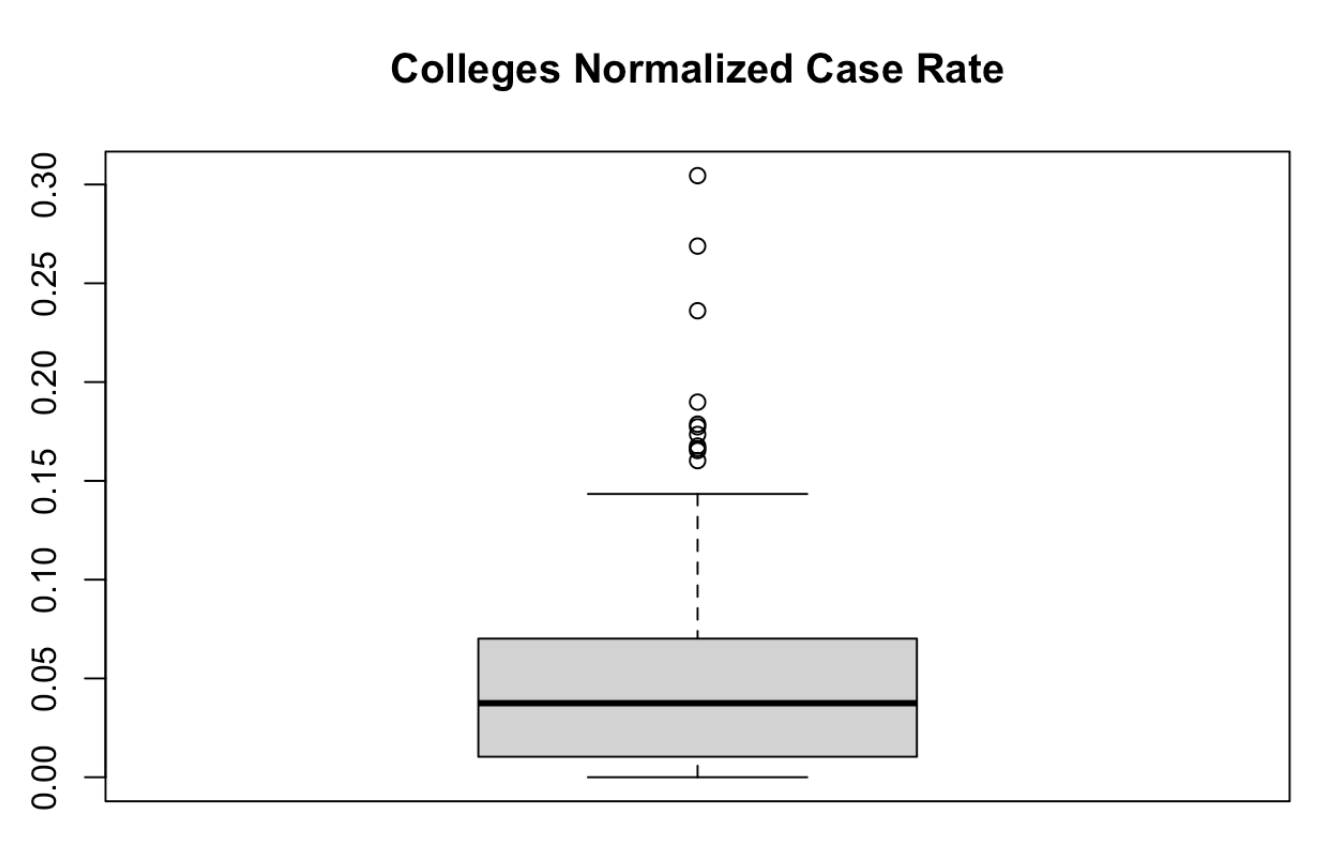
```
str(colleges_full)
```

```
## 'data.frame':    249 obs. of  19 variables:
##  $ X                 : int  1 2 3 4 5 6 9 12 15 19 ...
##  $ state             : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ county            : chr  "Lee" "Pike" "Jefferson" "Mobile" ...
##  $ city              : chr  "Auburn" "Troy" "Birmingham" "Mobile" ...
##  $ ipeds_id          : int  100858 102368 100663 102094 102553 104717 105330 104179 106467 106397 ...
##  $ college           : chr  "Auburn University" "Troy University" "University of Alabama at Birmingham" "Unive
rsity of South Alabama" ...
##  $ cases             : int  2742 855 3143 780 186 1935 2959 4365 704 3336 ...
##  $ rank2020          : int  101 285 193 338 348 5 94 37 442 129 ...
##  $ rank2015          : int  150 217 256 346 283 11 135 39 491 144 ...
##  $ enroll2020        : int  30440 16981 21923 14834 14644 90253 31066 44097 12101 27778 ...
##  $ enroll2015        : int  24864 20573 18568 15065 17363 55497 26594 40621 11369 25341 ...
##  $ NEVER             : num  0.053 0.06 0.049 0.077 0.042 0.023 0.039 0.023 0.082 0.034 ...
##  $ RARELY            : num  0.064 0.131 0.037 0.07 0.05 0.025 0.08 0.024 0.151 0.028 ...
##  $ SOMETIMES         : num  0.138 0.145 0.107 0.102 0.049 0.059 0.07 0.073 0.181 0.101 ...
##  $ FREQUENTLY        : num  0.183 0.333 0.179 0.244 0.196 0.158 0.11 0.121 0.285 0.186 ...
##  $ ALWAYS            : num  0.562 0.33 0.628 0.506 0.663 0.734 0.702 0.759 0.302 0.651 ...
##  $ winner            : chr  "Republican" "Republican" "Democrat" "Republican" ...
##  $ perc_enroll_change: num  0.2243 -0.1746 0.1807 -0.0153 -0.1566 ...
##  $ case_rate         : num  0.0901 0.0504 0.1434 0.0526 0.0127 ...
```

```
###### EDA
boxplot(colleges_full$cases, main = "Colleges Total Cases")
```


Colleges Total Cases

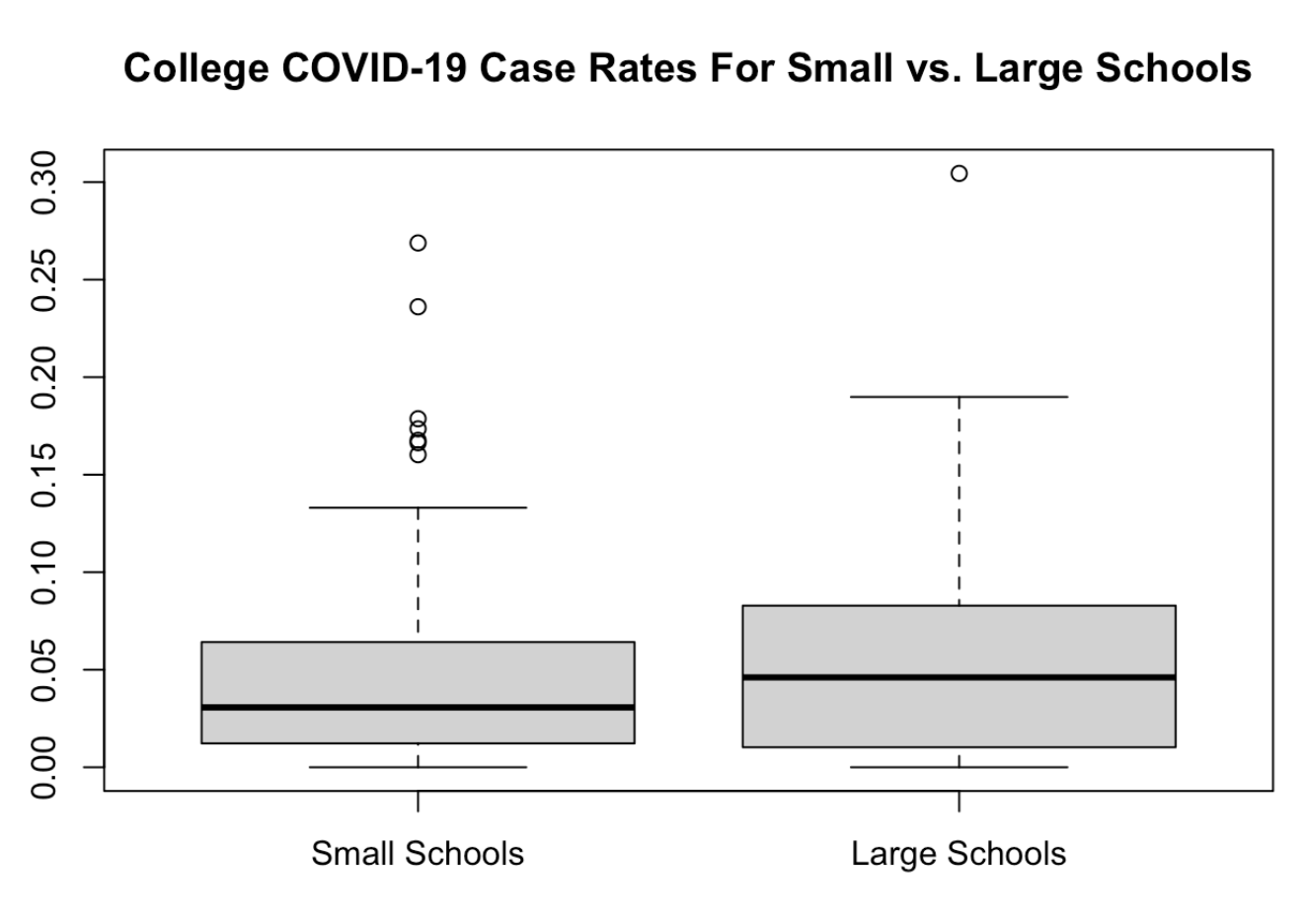
```
boxplot(colleges_full$case_rate,
        main = "Colleges Normalized Case Rate")
```


Colleges Normalized Case Rate

```
### Create Size Subgroups
summary(colleges_full$enroll2020)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11211   14029   18373   22283   26506  104068
```

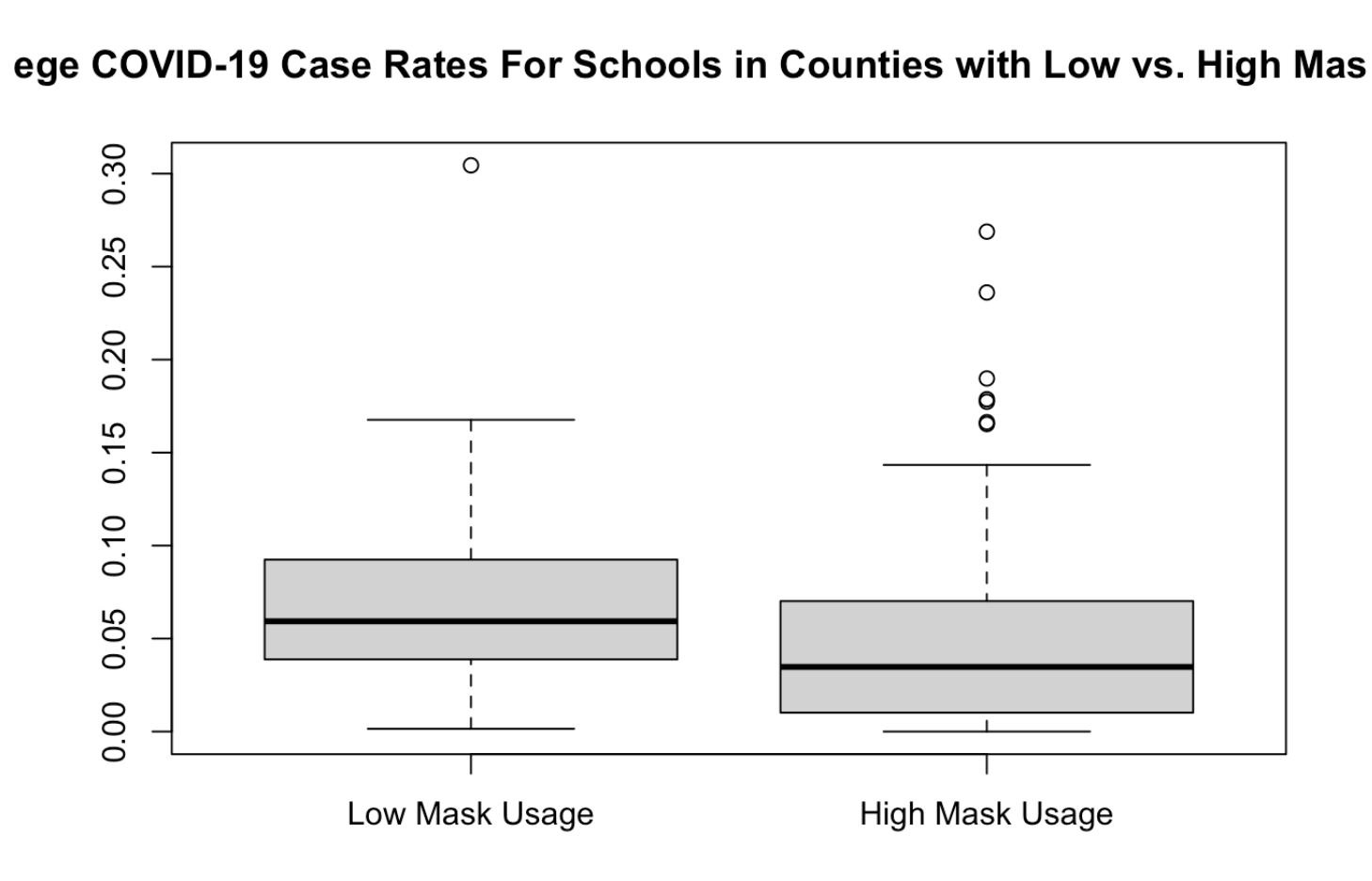
```
small_schools <- colleges_full$case_rate[colleges_full$enroll2020 <= 18000]
big_schools <- colleges_full$case_rate[colleges_full$enroll2020 > 18000]

boxplot(small_schools, big_schools,
        main = "College COVID-19 Case Rates For Small vs. Large Schools",
        names = c("Small Schools", "Large Schools"))
```


College COVID-19 Case Rates For Small vs. Large Schools

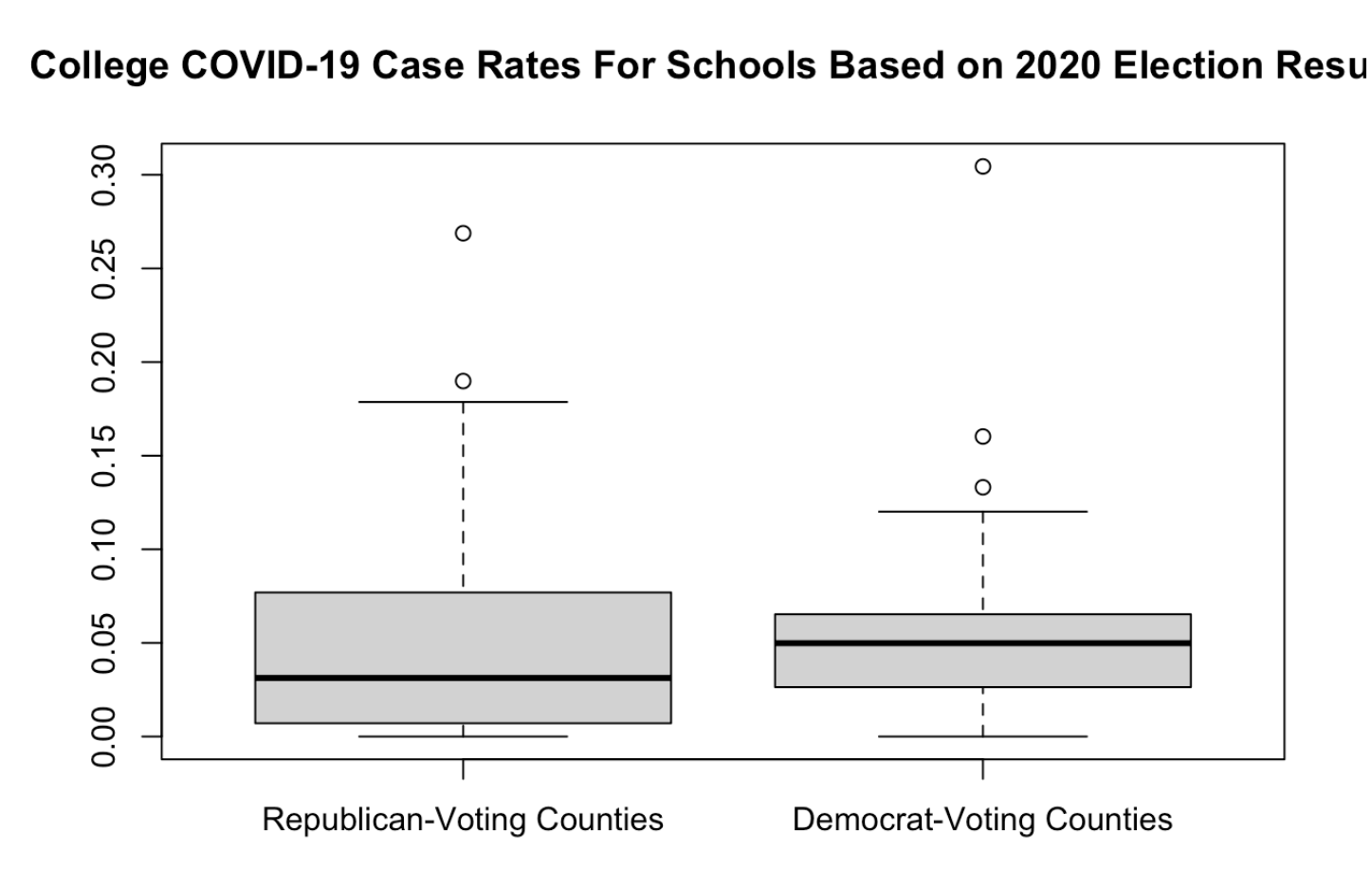
```
### Create Mask Wearing Subgroups
summary(colleges_full$ALWAYS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.2450  0.5675  0.6805  0.6507  0.7560  0.8550      27
```

```
high_mask_usage <- colleges_full$case_rate[colleges_full$ALWAYS > 0.5]
low_mask_usage <- colleges_full$case_rate[colleges_full$ALWAYS <= 0.5]
# Remove NAs
high_mask_usage<-high_mask_usage[!is.na(high_mask_usage)]
low_mask_usage<-low_mask_usage[!is.na(low_mask_usage)]


boxplot(low_mask_usage, high_mask_usage,
        main = "College COVID-19 Case Rates For Schools in Counties with Low vs. High Mask Usage",
        names = c("Low Mask Usage", "High Mask Usage"))
```


ege COVID-19 Case Rates For Schools in Counties with Low vs. High Mas

```
### Create Voting Subgroups
red_counties <- colleges_full$case_rate[colleges_full$winner == "Democrat"]
blue_counties <- colleges_full$case_rate[colleges_full$winner == "Republican"]
# Remove NAs
red_counties<-red_counties[!is.na(red_counties)]
blue_counties<-blue_counties[!is.na(blue_counties)]



boxplot(red_counties, blue_counties,
        main = "College COVID-19 Case Rates For Schools Based on 2020 Election Results",
        names = c("Republican-Voting Counties", "Democrat-Voting Counties"))
```


College COVID-19 Case Rates For Schools Based on 2020 Election Resu

```
######### Confidence Intervals
# Difference between small vs. large schools
school_size_bootstrap <- rep(NA,10000)

for (j in 1:10000) {
  boot.large <- mean(sample(big_schools, length(big_schools), replace = T))
  boot.small <- mean(sample(small_schools, length(small_schools), replace = T))
  school_size_bootstrap[j] <- boot.large - boot.small #the difference
}

head(school_size_bootstrap)
```

```
## [1] 0.013121659 0.003650055 0.003651378 0.004916144 0.008419205 0.015279614
```

```
mean(school_size_bootstrap) #bootstrap mean difference
```

```
## [1] 0.005595562
```

```
CI <- quantile(school_size_bootstrap, c(.05, .95))
CI
```

```
##           5%          95%
## -0.004424769  0.015686788
```

```
# Difference between schools in counties that wear masks and those that don't
mask_bootstrap <- rep(NA,100000)

for (j in 1:100000) {
  mask_high <- mean(sample(high_mask_usage, length(high_mask_usage), replace = T))
  mask_low <- mean(sample(low_mask_usage, length(low_mask_usage), replace = T))
  mask_bootstrap[j] <- mask_high - mask_low #the difference
}

head(mask_bootstrap)
```

```
## [1] -0.02382377 -0.04405964 -0.01660564 -0.02162277 -0.02479056 -0.02844865
```

```
mean(mask_bootstrap) #bootstrap mean difference
```

```
## [1] -0.02434536
```

```
CI_mask <- quantile(mask_bootstrap, c(.05, .95))
CI_mask
```

```
##           5%          95%
## -0.041403003 -0.008891409
```

```
# Difference between schools in counties that wear masks and those that don't
election_bootstrap <- rep(NA,100000)

for (j in 1:100000) {
  election_red <- mean(sample(red_counties, length(red_counties), replace = T))
  election_blue <- mean(sample(blue_counties, length(blue_counties), replace = T))
  election_bootstrap[j] <- election_red - election_blue #the difference
}

head(election_bootstrap)
```

```
## [1] -0.006414878 -0.006457495 -0.008835648 -0.016302977 -0.021318037
## [6] -0.004235977
```

```
mean(election_bootstrap) #bootstrap mean difference
```

```
## [1] -0.007115239
```

```
CI_election <- quantile(election_bootstrap, c(.05, .95))
CI_election
```

```
##           5%          95%
## -0.017876188  0.003197215
```

```
#########################
##### Prison Data

# Load in data
facilities <- read.csv("~/Desktop/Papers/Georgetown/Fall 2021/ANLY 511/Final Project/facilities_edited.csv")
head(facilities)
```

```
##   X                                                          facility_name
## 1 1                                            Alex City Work Release prison
## 2 2                            Alabama Therapeutic Education Facility prison
## 3 3                                               Bibb Correctional Facility
## 4 4 Birmingham Women's Community Based Facility and Community Work Center
## 5 5                                            Bullock Correctional Facility
## 6 6                                                            Camden prison
##                 facility_type    county facility_county_fips   state
## 1   Low-security work release     Coosa                 1037 Alabama
## 2 State rehabilitation center    Shelby                 1117 Alabama
## 3                State prison      Bibb                 1007 Alabama
## 4                State prison Jefferson                 1073 Alabama
## 5                State prison   Bullock                 1011 Alabama
## 6                State prison    Wilcox                 1131 Alabama
##   latest_inmate_population total_inmate_cases total_inmate_deaths
## 1                      188                 77                   0
## 2                      272                 11                   1
## 3                     1725                164                   3
## 4                      192                 17                   0
## 5                     1477                162                   5
## 6                       49                  5                   0
##   political_party county_population
## 1      Republican             10387
## 2      Republican            223024
## 3      Republican             22293
## 4        Democrat            674721
## 5        Democrat             10357
## 6        Democrat             10600
```

```
str(facilities)
```

```
## 'data.frame':    2553 obs. of  11 variables:
##  $ X                       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ facility_name           : chr  "Alex City Work Release prison" "Alabama Therapeutic Education Facility pris
on" "Bibb Correctional Facility" "Birmingham Women's Community Based Facility and Community Work Center" ...
##  $ facility_type           : chr  "Low-security work release" "State rehabilitation center" "State prison" "St
ate prison" ...
##  $ county                  : chr  "Coosa" "Shelby" "Bibb" "Jefferson" ...
##  $ facility_county_fips    : int  1037 1117 1007 1073 1011 1131 1121 1073 1051 1005 ...
##  $ state                   : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ latest_inmate_population: int  188 272 1725 192 1477 49 322 1508 100 1181 ...
##  $ total_inmate_cases      : int  77 11 164 17 162 5 5 45 121 72 ...
##  $ total_inmate_deaths     : int  0 1 3 0 5 0 0 2 0 2 ...
##  $ political_party         : chr  "Republican" "Republican" "Republican" "Democrat" ...
##  $ county_population       : int  10387 223024 22293 674721 10357 10600 82149 674721 87977 25223 ...
```

```
facilities$facility_type = as.factor(facilities$facility_type)
facilities <- facilities %>%
  mutate(case_rate = total_inmate_cases/latest_inmate_population) %>%
  filter(case_rate < 1)

### Confidence Intervals
levels(facilities$facility_type)
```

```
##  [1] "Detention center"            "Federal halfway house"
##  [3] "Federal prison"              "Jail"
##  [5] "Juvenile detention at jail"  "Low-security work release"
##  [7] "Reservation jail"            "State facility"
##  [9] "State halfway house"         "State juvenile detention"
## [11] "State prison"                "State rehabilitation center"
## [13] "State work camp"
```

```
# Break up data into subgroups
state_prisons <- facilities$case_rate[facilities$facility_type == c("State halfway house", "State work camp",
                                               "State juvenile detention", "State prison", "State facility",
                                               "State rehabilitation center")]
federal_prisons <- facilities$case_rate[facilities$facility_type == c("Detention center", "Juvenile detention at
jail",
                                                 "Federal halfway house", "Low-security work release", "Federal p
rison",
                                                 "Reservation jail", "Jail")]
```

```
## Warning in `==.default`(facilities$facility_type, c("Detention center", : longer
## object length is not a multiple of shorter object length
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
state_prisons<-state_prisons[!is.na(state_prisons)]
federal_prisons<-federal_prisons[!is.na(federal_prisons)]



small_prisons <- facilities$case_rate[facilities$latest_inmate_population < 580]
big_prisons <- facilities$case_rate[facilities$latest_inmate_population >= 580]
# Remove NAs
small_prisons<-small_prisons[!is.na(small_prisons)]
big_prisons<-big_prisons[!is.na(big_prisons)]



red_prisons <- facilities$case_rate[facilities$political_party == "Republican"]
blue_prisons <- facilities$case_rate[facilities$political_party == "Democrat"]
red_prisons<-red_prisons[!is.na(red_prisons)]
blue_prisons<-blue_prisons[!is.na(blue_prisons)]

# 1. Mean case numbers between different types of prisons
prison_type_bootstrap <- rep(NA,1000)

for (j in 1:1000) {
  boot.state <- mean(sample(state_prisons, length(state_prisons), replace = T))
  boot.federal <- mean(sample(federal_prisons, length(federal_prisons), replace = T))
  prison_type_bootstrap[j] <- boot.state - boot.federal #the difference
}

head(prison_type_bootstrap)
```

```
## [1] -0.098180539 -0.031825082 -0.010370224  0.002841252 -0.049518496
## [6] -0.029431810
```

```
mean(prison_type_bootstrap) #bootstrap mean difference
```

```
## [1] -0.0160858
```

```
prison_type_bootstrap<-prison_type_bootstrap[!is.na(prison_type_bootstrap)]

CI_prison1 <- quantile(prison_type_bootstrap, c(.05, .95))
CI_prison1
```

```
##          5%         95%
## -0.08014270  0.04672863
```

```
# 2. Mean case numbers between different sized prisons
prison_type_bootstrap2 <- rep(NA,1000)

for (j in 1:1000) {
  boot.small <- mean(sample(small_prisons, length(small_prisons), replace = T))
  boot.big <- mean(sample(big_prisons, length(big_prisons), replace = T))
  prison_type_bootstrap2[j] <- boot.big - boot.small #the difference
}

head(prison_type_bootstrap2)
```

```
## [1] 0.03812381 0.04982210 0.07058573 0.06497707 0.03413541 0.05169557
```

```
mean(prison_type_bootstrap2) #bootstrap mean difference
```

```
## [1] 0.05600088
```

```
CI_prison2 <- quantile(prison_type_bootstrap2, c(.05, .95))
CI_prison2
```

```
##         5%        95%
## 0.03392240 0.07758713
```

```
# 3. Mean case numbers between prisons in Red vs. Blue states - facilities.csv
prison_type_bootstrap3 <- rep(NA,1000)

for (j in 1:1000) {
  boot.red <- mean(sample(red_prisons, length(red_prisons), replace = T))
  boot.blue <- mean(sample(blue_prisons, length(blue_prisons), replace = T))
  prison_type_bootstrap3[j] <- boot.red - boot.blue #the difference
}

head(prison_type_bootstrap3)
```

```
## [1] 0.05228466 0.04810271 0.05397790 0.03854286 0.02015913 0.02123528
```

```
mean(prison_type_bootstrap3) #bootstrap mean difference
```

```
## [1] 0.04140843
```

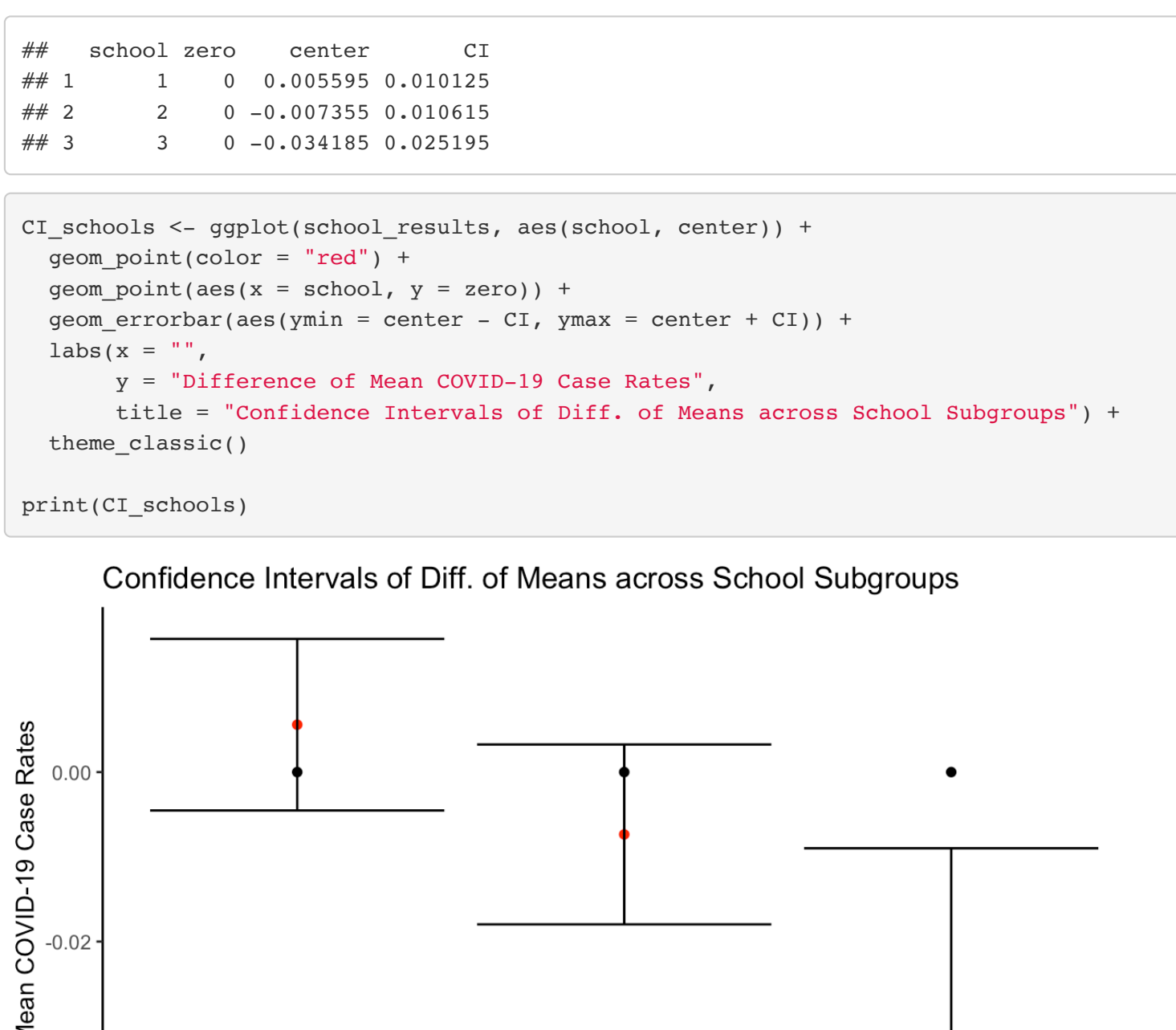
```
CI_prison3 <- quantile(prison_type_bootstrap3, c(.05, .95))
CI_prison3
```

```
##         5%        95%
## 0.01782058 0.06313119
```

```
#########################
## Visualize

school_results <- data.frame(school = c(1, 2, 3),
                   zero    = c(0, 0, 0),
                   center   = c(0.005595, -0.007355, -0.034185),
                   CI       = c(0.010125, 0.010615, 0.025195))
print(school_results)
```

```
##   school zero    center       CI
## 1      1    0  0.005595 0.010125
## 2      2    0 -0.007355 0.010615
## 3      3    0 -0.034185 0.025195
```

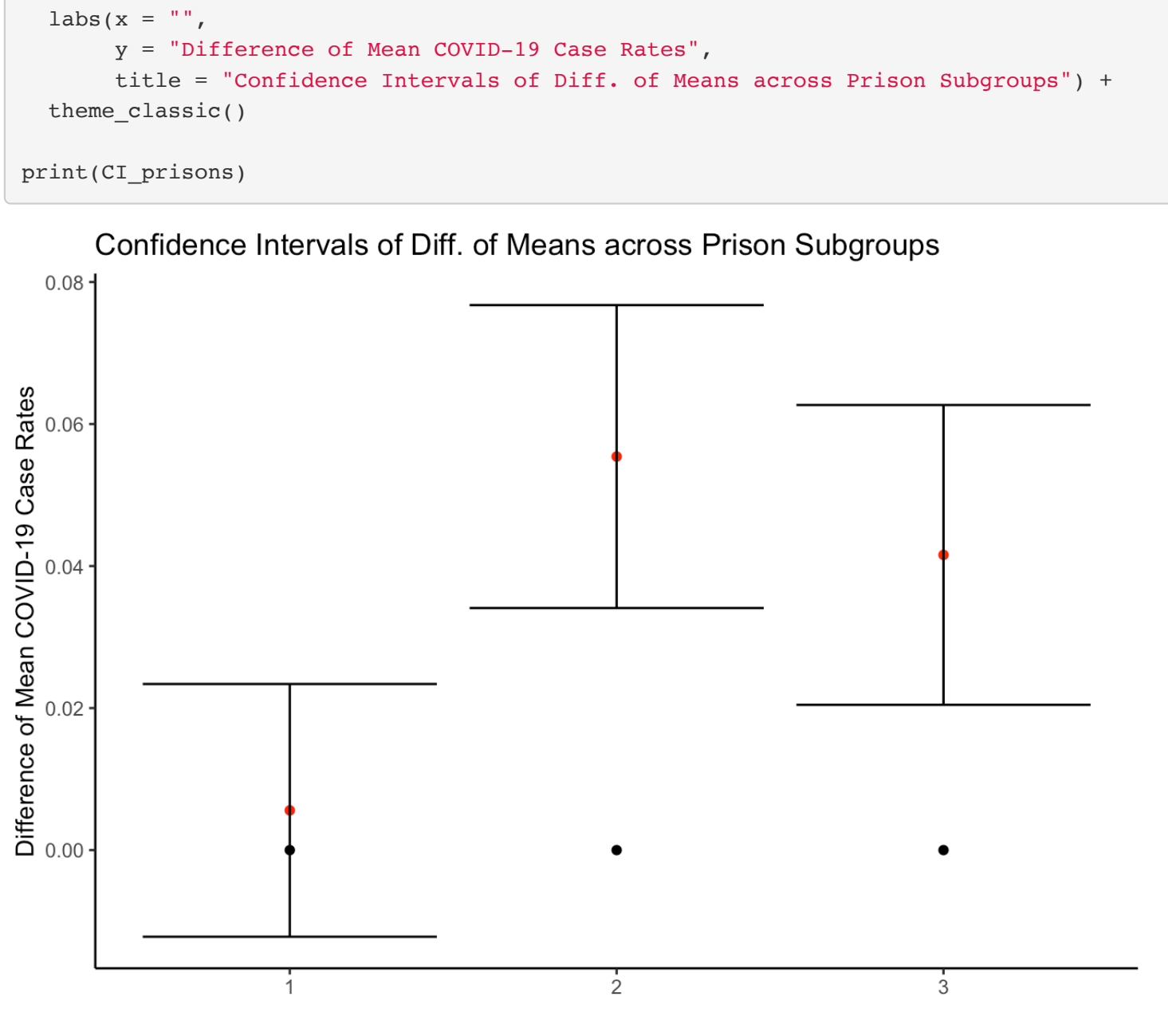
```
CI_schools <- ggplot(school_results, aes(school, center)) +
  geom_point(color = "red") +
  geom_point(aes(x = school, y = zero)) +
  geom_errorbar(aes(ymin = center - CI, ymax = center + CI)) +
  labs(x = "",
       y = "Difference of Mean COVID-19 Case Rates",
       title = "Confidence Intervals of Diff. of Means across School Subgroups") +
  theme_classic()

print(CI_schools)
```


Confidence Intervals of Diff. of Means across School Subgroups

```
# Prison data
prison_results <- data.frame(tests = c(1, 2, 3),
                             zero    = c(0, 0, 0),
                             center   = c(0.005595, 0.05542, 0.04157),
                             CI       = c(-0.017795, 0.02133, 0.02111))
print(school_results)
```

```
##   school zero    center       CI
## 1      1    0  0.005595 0.010125
## 2      2    0 -0.007355 0.010615
## 3      3    0 -0.034185 0.025195
```

```
CI_prisons <- ggplot(prison_results, aes(tests, center)) +
  geom_point(color = "red") +
  geom_point(aes(x = tests, y = zero)) +
  geom_errorbar(aes(ymin = center - CI, ymax = center + CI)) +
  labs(x = "",
       y = "Difference of Mean COVID-19 Case Rates",
       title = "Confidence Intervals of Diff. of Means across Prison Subgroups") +
  theme_classic()

print(CI_prisons)
```


Confidence Intervals of Diff. of Means across Prison Subgroups